



Development and application of QSAR models for mechanisms related to endocrine disruption.

Abildgaard Rosenberg, Sine

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):

Abildgaard Rosenberg, S. (2017). *Development and application of QSAR models for mechanisms related to endocrine disruption*. National Food Institute, Technical University of Denmark.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Development and application of QSAR models for mechanisms related to endocrine disruption

PhD thesis

Sine Abildgaard Rosenberg

Division for Diet, Disease Prevention and Toxicology

National Food Institute

Technical University of Denmark

April 2017

Thesis Title

Development and application of QSAR models for mechanisms related to endocrine disruption

Author

Sine Abildgaard Rosenberg

Supervisors

Senior Officer Eva Bay Wedebye

Senior Scientist Nikolai Georgiev Nikolov

Senior Scientist Marianne Dybdahl

Professor Anne Marie Vinggaard

Division for Diet, Disease Prevention and Toxicology

National Food Institute

Technical University of Denmark

Evaluation Committee

Julie Boberg, Senior Scientist (Chair), Technical University of Denmark, Denmark

Mark Timothy David Cronin, Professor, Liverpool John Moores University, England

Stefan Theodor Kramer, Professor, Johannes Gutenberg University of Mainz, Germany

Funding

This project was financially supported by the Danish 3R center (one third) and the Technical University of Denmark (two thirds).

Copyright

National Food Institute, Technical University of Denmark

Photo

Sine A. Rosenberg (right, word cloud; left, modified Leadscope® table).

ISBN

978-87-93565-04-3

This PhD Thesis is available at

www.food.dtu.dk

National Food Institute

Technical University of Denmark

Kemitorvet

Building 202

2800 Kgs. Lyngby

Tel: +45 35 88 70 00

Fax: +45 35 88 70 01

Preface

The work included in this thesis was carried out in the period from December 2013 to April 2017 at the National Food Institute, Technical University of Denmark, and the National Center for Computational Toxicology, U.S. Environmental Protection Agency, North Carolina. During my work I was supervised by Eva B. Wedebye, Nikolai G. Nikolov, Marianne Dybdahl and Anne Marie Vinggaard from the National Food Institute. All my supervisors are greatly acknowledged for their continuous support and guidance during my PhD project. I would also like to thank the National Food Institute, who gave financial support to 2/3 of the project, as well as the Danish 3R Center for a grant for the last 1/3 of the project.

I would like to express my greatest gratitude to all my colleagues at the National Food Institute for a very supportive and caring working environment. A special thanks to my office mates Camilla Schwartz, Hanna Johansson, Katrine Frederiksen and Karin Lauschke for contributing to an amazing office atmosphere and supporting me during the last months of my PhD. Hanna, who has followed me from the beginning of my PhD and who is now a close friend deserves extra thanks. Dr. Richard Judson, my mentor at the National Center of Computational Toxicology, and his colleagues are thanked for their very warm welcome – you made my 5 months in North Carolina to an unforgettable and professionally as well as personally developing time. Last, but not least, I would like to thank my friends, family and partner for always being there.

Søborg, April 2017

Sine Rosenberg

Table of Contents

Summary	i
Dansk Resumé	iii
List of Papers and Manuscripts	v
List of Abbreviations	i
PART I - Introduction	1
1.1 Motivation and Scope of the Project	3
1.2 Organization of the Thesis	4
PART II - Background	7
2.1 The Endocrine System and Endocrine Disrupting Chemicals	9
2.1.1 The Endocrine System	9
2.1.2 Endocrine Disrupting Chemicals	12
2.2 Quantitative Structure-Activity Relationship Models	21
2.2.1 QSAR Development	21
2.2.2 The OECD Principles for Validation of QSAR Models	32
2.2.3 The Danish (Q)SAR Database	34
2.2.4 Application of QSAR	35
2.3 Regulatory Toxicology	47
2.3.1 A Paradigm Shift in Toxicology	47
2.3.2 ToxCast and Tox21 Programs	49
2.3.3 Adverse Outcome Pathways	51
2.3.4 Integrated Approaches to Testing and Assessment	52
2.3.5 Registration, Evaluation and Authorisation of Chemicals	53
PART III - Projects	63
3.1 QSAR Models for TPO Inhibition <i>In Vitro</i>	65
3.1.1 Manuscript in Preparation	65
3.2 QSAR Models for PXR Interaction and CYP3A4 Induction <i>In Vitro</i>	91
3.2.1 Published Paper	91
3.3 QSAR Models for AhR Activation <i>In Vitro</i>	101
3.3.1 Study Report	101
3.4 The Collaborative Estrogen Receptor Activity Prediction Project	115
3.4.1 Introduction	115
3.4.2 My Contributions to CERAPP	115
3.4.3 Published paper	117
3.4.4 My Further Remarks to CERAPP	128
3.4.5 Conclusions	128

Part IV - In Closing.....	131
4.1 Overview	133
4.2 Discussion	136
4.2.1 Collection, Curation and Preparation of the Applied Datasets	136
4.2.2 QSAR Development	137
4.2.3 Limitations of the Developed QSAR Models.....	138
4.2.4 Using the Developed QSAR Models.....	139
4.3 Concluding Remarks.....	140
4.4 Perspectives	141
Appendix.....	145

Summary

Humans are daily exposed to a wide variety of man-made chemicals through food, consumer products, water, air inhalation etc. For the main part of these chemicals no or only very limited information is available on their potential to cause endocrine disruption. Traditionally such information has been derived from animal studies, which are time-consuming, expensive and subject to ethical issues. For these reasons alternative methods such as cell culture studies and non-testing approaches such as quantitative structure-activity relationships (QSARs) are of high value as they can provide information on the mode of action of chemicals in a faster and cheaper way. The main purpose in this PhD project was to develop QSAR models for mechanisms related to endocrine disruption and apply the models to predict 10,000s of chemicals to which humans are potentially exposed.

The first part of the thesis is a background section, comprising 1) an introduction to the endocrine system with a focus on thyroid hormones (THs) and their essential function in neurodevelopment as well as a description of how chemicals may interfere with endocrine mechanisms and cause adverse effects, 2) an introduction to the applied methods to develop QSARs, and 3) an introduction to regulatory toxicology including the acceptance of predictions from QSARs under the European chemicals regulation, REACH. Following the background section, the four projects of the thesis are described. The first three projects focus on the development of QSARs for mechanisms that can affect TH levels: Thyroperoxidase (TPO) inhibition, Pregnane X receptor (PXR) activation, and Aryl hydrocarbon receptor (AhR) activation. TPO is an enzyme essential in the synthesis of THs, and both PXR and AhR are important regulators of enzymes involved in the turnover of THs and other hormones. The fourth project was part of a large international QSAR collaboration, CERAPP, in which a QSAR model for estrogen receptor (ER) agonism was developed, and used to predict 32,197 CERAPP chemicals. All models in the four projects were validated to assess how good they are at making correct predictions, and they all showed good predictive performance. The QSAR models were used to predict 72,524 REACH substances, and they were able to predict between 38,114 to 53,433 of these substances.

To conclude, the QSAR models developed in this PhD project can provide important information on the 10,000s of chemicals in our surroundings. The predictions can for example be used for prioritizing chemicals for further evaluation, aid in chemical assessments, grouping approaches, and drug development as well as in the generation of new hypotheses on mode of actions in adverse health outcomes.

Dansk Resumé

Mennesker udsættes dagligt for mange forskellige kemikalier fra fx madvarer, personlig pleje produkter, vand og luften. For størstedelen af disse kemikalier er der ingen eller kun meget begrænset viden om deres potentielle hormonforstyrrende effekter. Traditionelt har man indsamlet denne information fra dyreforsøg, men de er tidskrævende, dyre og etisk problematiske. Alternative metoder såsom celleforsøg og computermødelles som f.eks. quantitative structure-activity relationships (QSARs) kan bruges til på en hurtigere og billigere måde at forstå kemikalernes virkningsmekanismer. Hovedformålet med dette PhD projekt var at udvikle QSAR modeller for mekanismer i hormonsystemet, og benytte disse modeller til at screene 10.000'er af kemikalier, som mennesker potentielt udsættes for.

Første del af afhandlingen består af et baggrundsafsnit, der 1) introducerer hormonsystemet med fokus på thyreoideahormoner (TH'er), som bl.a. er essentielle i udviklingen af hjernen, samt beskriver, hvordan kemikalier kan påvirke mekanismer hormonsystemet og derigennem forårsage sundhedsskadelige effekter, 2) introducerer de metoder der anvendes i udviklingen af QSAR modeller, og 3) introducerer den regulatoriske toksikologi, og hvordan QSAR forudsigelser bl.a. kan benyttes i den Europæiske kemikalielovgivning, REACH.

I næste del beskrives afhandlingens fire projekter. I de første tre projekter blev der udviklet QSAR modeller for mekanismer, som påvirker TH niveauet: Thyroperoxidase (TPO) hæmning, Pregnane X receptor (PXR) aktivering, og Aryl hydrocarbon receptor (AhR) aktivering. TPO er et vigtigt enzym i syntesen af TH'er, og både PXR og AhR er vigtige i reguleringen af enzymer involveret i omsætningen af TH'er og andre hormoner. Det fjerde projekt var en del af et stort internationalt QSAR samarbejde, CERAPP. Hertil blev der udviklet en QSAR model for østrogen receptor aktivering, en vigtig mekanisme for hormonforstyrrende kemikalier, og modellen blev brugt til at forudsige 32.197 CERAPP kemikalier. Alle modellerne blev valideret for at vurdere deres evne til at lave korrekte forudsigelser, og de viste alle høje nøjagtigheder. Modellerne blev efterfølgende bl.a. brugt til at forudsige 72.524 REACH stoffer, og de kunne forudsige mellem 38.114 og 53.433 af stofferne.

De udviklede QSAR modeller kan bidrage med værdifuld information om de 10.000-vis af kemikalier i vores omgivelser. Forudsigelserne kan bl.a. bruges til at prioritere kemikalier til yderligere toksikologisk vurdering, samt blive brugt i evalueringen og grupperingen af kemikalier, i udviklingen af lægemidler og i opstillingen af nye hypoteser om underliggende virkningsmekanismer i sundhedsskadelige effekter.

List of Papers and Manuscripts

Published Papers

S.A. Rosenberg, M. Xia, R. Huang, N.G. Nikolov, E.B. Wedebye, M. Dybdahl, QSAR development and profiling of 72,524 REACH substances for PXR activation and CYP3A4 induction, *Comput. Toxicol.* 1 (2017) 39–48. doi:10.1016/j.comtox.2017.01.001.

K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A.M. Richard, C.M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E.B. Wedebye, F. Grisoni, G.F. Mangiatordi, G.M. Incisivo, H. Hong, H.W. Ng, I. V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N.G. Nikolov, O. Nicolotti, P.L. Andersson, Q. Zang, R. Politi, R.D. Beger, R. Todeschini, R. Huang, S. Farag, **S.A. Rosenberg**, S. Slavov, X. Hu, R.S. Judson, CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ. Health Perspect.* 124 (2016) 1023–1033. doi:10.1289/ehp.1510267.

Manuscript in Preparation

S.A. Rosenberg, E.D. Watt, R.S. Judson, S.O. Simmons, K. Paul Friedman, M. Dybdahl, N.G. Nikolov, E.B. Wedebye, QSAR Models for Thyroperoxidase Inhibition and Screening of U.S. and EU Chemical Inventories.

Study Report

S.A. Rosenberg, M. Dybdahl, E.B. Wedebye, N.G. Nikolov, A pilot study to explore the effect of rational selection of training set inactives on model predictive performance and coverage using a large imbalanced AhR activation dataset.

List of Abbreviations

AD	Applicability domain	HPT	Hypothalamus-pituitary-thyroid
ADHD	Attention-deficit/hyperactivity disorder	HTS	High-throughput screening
AhR	Aryl hydrocarbon receptor	IATA	Integrated approaches to testing and assessment
ANN	Artificial neural network	ICH	International Council for Harmonisation
AO	Adverse outcome	InCHI	IUPAC International Chemical Identifier
AOP	Adverse outcome pathway	IPCS	International Programme on Chemical Safety
ASD	Autism spectrum disorders	IRD	Inner ring deiodinase
CAR	Constitutive androstane receptor	ITS	Integrated testing strategies
CART	Classification and regression trees	JRC	Joint Research Center
CERAPP	Collaborative Estrogen Receptor Activity Prediction Project	KE	Key event
CLP	Classification, labelling and packaging	<i>k</i> NN	<i>k</i> -nearest neighbors
CMR	Carcinogenic , mutagenic or toxic to reproduction	LBVS	Ligand-based virtual screening
DIT	Diiodotyrosine	LDA	Linear discriminant analysis
DNT	Developmental neurotoxicity	LMO	Leave-many -out
DTU	Technical University of Denmark	LOO	Leave-one-out
EC	European Commission	LPDM	Leadscope Predictive Data Miner
ECHA	European Chemicals Agency	MCT-8	Monocarboxylate transporter-8
EDC	Endocrine disrupting chemicals	MIE	Molecular initiating event
EDSP	Endocrine Disruptor Screening Program	MIT	Monoiodotyrosine
EINECS	European Inventory of Existing Commercial Chemical Substances	MLR	Multiple linear regression
EOGRTS	Extended one-generation reproductive toxicity study	MoA	Mode-of-action
EPA	Environmental Protection Agency	NB	Naïve bayes
ER	Estrogen receptor	NCATS	National Center for Advancing Translational Sciences
ERDC	Engineer Research & Development Center	NCBI	National Center for Biotechnology Information
FDA	Food and Drug Administration	NCCT	National Center for Computational Toxicology
FN	False negative	NCGS	NCATS Chemical Genomics Center
FP	False positive	NIEHS	National Institute of Environmental Health Sciences
GA	Genetic algorithm	NIH	National Institutes of Health
		NIS	Na ⁺ /I ⁻ symporter
		NR	Nuclear receptor
		NRC	National Research Council

NTP	National Toxicology Program	WHO	World Health Organization
OATP1c1	Organic anion transporter protein 1c1	WoE	Weight of evidence
OECD	Organisation for Economic Co-operation and Development		
ORD	Outer-ring deiodinase		
PCA	Principal component analysis		
PLR	Partial logistic regression		
PLS	Partial least squares		
PRESS	Predicted residual error sum of square		
PRS	Pre-registered substances		
PXR	Pregnane X receptor		
QSAR	Quantitative structure-activity relationship		
QMRF	QSAR model reporting format		
QPRF	QSAR prediction reporting format		
RF	Random forest		
rT3	Reverse triiodothyronine		
SAR	Structure-activity relationship		
SD	Standard deviation		
SMILES	Simplified molecular input line entry system		
SULT	Sulfotransferases		
SVM	Support vector machines		
T3	Triiodothyronine		
T4	Thyroxine		
TBG	Thyroxine binding globulin		
TDC	Thyroid disrupting chemical		
Tg	Thyroglobulin		
TH	Thyroid hormone		
TSH	Thyroid stimulating hormone		
TN	True negative		
TP	True positive		
TPO	Thyroperoxidase		
TR	Thyroid hormone receptor		
TRE	Thyroid hormone response elements		
TRH	Thyroid releasing hormone		
TTR	Transthyretin		
UGT	UDP-glucuronosyltransferases		

PART I - Introduction

1.1 Motivation and Scope of the Project

Humans are continuously exposed to a wide variety of man-made chemicals through for example food, water, consumer products such as cosmetics and house-cleaning products, pharmaceuticals, and air inhalation [1–4]. These chemicals have the potential to interfere with normal physiological systems of living organisms and, if the interferences are left uncompensated, adverse health effects may develop. Evidence from epidemiological studies indicates that chemical exposure is involved in a number of adverse human health effects such as cancer, reduced reproductive health and learning disabilities [5–11]. Some of these adverse outcomes are likely the result of chemical interference with molecular mechanisms of the endocrine system such as interaction with hormone receptors and/or altered synthesis, degradation or transport of natural hormones [8,12]. This has led to an increased focus on identifying chemicals with endocrine modulating properties, i.e. so-called endocrine disrupting chemicals, and screening for a battery of such properties has been included in programs and legislations within both EU and US [4,13,14].

Traditional toxicology testing consists of exposing laboratory animals, typically rats or mice, to a chemical and looking for adverse effects at whole animal, tissue and/or cellular level. Animal tests are time-consuming, expensive, subject to ethical issues, and their results can be difficult to extrapolate to humans [15–18]. Due to these challenges/limitations with animal toxicity tests and the ongoing need to gather toxicity information on the many thousands of chemicals in commerce, a paradigm shift in toxicity testing have been proposed, often referred to as Toxicity Testing in the 21st Century [19,20]. Here the use of alternative methods such as *in vitro* and *in silico* to aid in chemical safety assessment is presented [19–22].

In this PhD project, the *in silico* method Quantitative Structure-Activity Relationship (QSAR) modeling was applied on a number of molecular mechanisms within the endocrine system, most of which are molecular initiating events (MIEs) in established adverse outcome pathways (AOPs) of thyroid-related adverse outcomes [23–26]. The developed models underwent thorough validations according to regulatory recommendations [27] and were then used for screening of large chemical inventories containing man-made chemicals.

The main hypothesis of this PhD project is:

QSAR models for selected molecular mechanisms of thyroid-related AOPs can expand the knowledge derived from experimental data and aid in human health safety evaluation of chemicals.

To investigate this hypothesis, the following questions have been sought answered:

- Can highly predictive and robust global QSAR models for MIEs in relevant AOPs be developed?
- If so, can such QSAR models trained on 1,000s of structurally diverse chemicals, provide reliable predictions and hereby extend the use of information from tested chemicals to 10,000s of man-made untested chemicals?

1.2 Organization of the Thesis

The thesis is organized into four parts. Part I gives an introduction to the motivation for the PhD project, its scope, hypothesis and organization. In Part II a general background on the endocrine system and related toxicology with focus on the thyroid system is given followed by an outline on the concept of QSAR models and their applications, and finally an introduction to regulatory toxicology. The background sections in Part II are not exhaustive and more information on the different topics may be found in the published literature. Part III contains separate chapters describing each of the four projects of this thesis. Accepted papers, submitted manuscripts or study reports from each of the projects are included in the respective chapters. The final Part IV consists of a brief overview, a summarizing discussion and conclusion as well as further research perspectives.

References

- [1] K.L. Dionisio, A.M. Frame, M.-R. Goldsmith, J.F. Wambaugh, A. Liddell, T. Cathey, D. Smith, J. Vail, A.S. Ernstoff, P. Fantke, O. Jolliet, R.S. Judson, Exploring consumer exposure pathways and patterns of use for chemicals in the environment, *Toxicol. Reports*. 2 (2015) 228–237. doi:10.1016/j.toxrep.2014.12.009.
- [2] P.P. Egeghy, R. Judson, S. Gangwal, S. Mosher, D. Smith, J. Vail, E.A. Cohen Hubal, The exposure data landscape for manufactured chemicals, *Sci. Total Environ.* 414 (2012) 159–166. doi:10.1016/j.scitotenv.2011.10.046.
- [3] M.-R. Goldsmith, C.M. Grulke, R.D. Brooks, T.R. Transue, Y.M. Tan, A. Frame, P.P. Egeghy, R. Edwards, D.T. Chang, R. Tornero-Velez, K. Isaacs, A. Wang, J. Johnson, K. Holm, M. Reich, J. Mitchell, D.A. Vallero, L. Phillips, M. Phillips, J.F. Wambaugh, R.S. Judson, T.J. Buckley, C.C. Dary, Development of a consumer product ingredient database for chemical exposure screening and prioritization, *Food Chem. Toxicol.* 65 (2014) 269–279. doi:10.1016/j.fct.2013.12.029.
- [4] R. Judson, A. Richard, D.J. Dix, K. Houck, M. Martin, R. Kavlock, V. Dellarco, T. Henry, T. Holderman, P. Sayre, S. Tan, T. Carpenter, E. Smith, The Toxicity Data Landscape for Environmental Chemicals, *Environ. Health Perspect.* 117 (2009) 685–695. doi:10.1289/ehp.0800168.
- [5] Å. Bergman, J.J. Heindal, S. Jobling, K.A. Kidd, R.T. Zoeller, State of the science of endocrine disrupting chemicals 2012, World Health Organization and United Nations Environment Programme, 2013. http://apps.who.int/iris/bitstream/10665/78102/1/WHO_HSE_PHE_IHE_2013.1_eng.pdf (accessed March 13, 2017).
- [6] A. Blair, N. Kazerouni, Reactive chemicals and cancer, *Cancer Causes Control*. 8 (1997) 473–490. doi:10.1023/A:1018417623867.
- [7] P. Grandjean, P.J. Landrigan, Neurobehavioural effects of developmental toxicity, *Lancet Neuro.* 13 (2014) 330–338. doi:10.1016/S1474-4422(13)70278-3.
- [8] P.T.C. Harrison, P. Holmes, C.D.N. Humfrey, Reproductive health in humans and wildlife: are adverse trends associated with environmental chemical exposure?, *Sci. Total Environ.* 205 (1997) 97–106. doi:10.1016/S0048-9697(97)00212-X.
- [9] S.H. Swan, K.M. Main, F. Liu, S.L. Stewart, R.L. Kruse, A.M. Calafat, C.S. Mao, J.B. Redmon, C.L. Ternand, S. Sullivan, J.L. Teague, E.Z. Drobni, B.S. Carter, D. Kelly, T.M. Simmons, C. Wang, L. Lumbreras, S. Villanueva, M. Diaz-Romero, M.B. Lomeli, E. Otero-Salazar, C. Hobel, B. Brock, C. Kwong, A. Muehlen, A. Sparks, A. Wolf, J. Whitham, M. Hatterman-Zogg, M. Maifeld, Decrease in anogenital distance among male infants with prenatal phthalate exposure, *Environ. Health Perspect.* 113 (2005) 1056–1061. doi:10.1289/ehp.8100.
- [10] C. Wohlfahrt-Veje, H.R. Andersen, I.M. Schmidt, L. Aksglaede, K. Sørensen, A. Juul, T.K. Jensen, P. Grandjean, N.E. Skakkebaek, K.M. Main, Early breast development in girls after prenatal exposure to non-persistent pesticides, *Int. J. Androl.* 35 (2012) 273–282. doi:10.1111/j.1365-2605.2011.01244.x.
- [11] WHO, Endocrine disruptors and child health: Possible developmental early effects of endocrine disruptors on child health, (2012). http://apps.who.int/iris/bitstream/10665/75342/1/9789241503761_eng.pdf (accessed March 13, 2017).
- [12] E.S. Tien, M. Negishi, Nuclear receptors CAR and PXR in the regulation of hepatic metabolism, *Xenobiotica*. 36 (2006) 1152–1163. doi:10.1080/00498250600861827.

- [13] EDSP21 Work Plan, The Incorporation of In Silico Models and In Vitro High Throughput Assays in the Endocrine Disruptor Screening Program (EDSP) for Prioritization and Screening, (2011). https://www.epa.gov/sites/production/files/2015-07/documents/edsp21_work_plan_summary_overview_final.pdf (accessed March 13, 2017).
- [14] REACH, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), (2006). <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02006R1907-20161011&from=EN>.
- [15] D. Fourches, J.C. Barnes, N.C. Day, P. Bradley, J.Z. Reed, A. Tropsha, Cheminformatics Analysis of Assertions Mined from Literature That Describe Drug-Induced Liver Injury in Different Species, *Chem. Res. Toxicol.* 23 (2010) 171–183. doi:10.1021/tx900326k.
- [16] M.I. Martić-Kehl, R. Schibli, P.A. Schubiger, Can animal data predict human outcome? Problems and pitfalls of translational animal research, *Eur. J. Nucl. Med. Mol. Imaging.* 39 (2012) 1492–1496. doi:10.1007/s00259-012-2175-z.
- [17] H. Olson, G. Betton, D. Robinson, K. Thomas, A. Monro, G. Kolaja, P. Lilly, J. Sanders, G. Sipes, W. Bracken, M. Dorato, K. Van Deun, P. Smith, B. Berger, A. Heller, Concordance of the Toxicity of Pharmaceuticals in Humans and in Animals, *Regul. Toxicol. Pharmacol.* 32 (2000) 56–67. doi:10.1006/rtph.2000.1399.
- [18] K. Stanton, F.H. Kruszewski, Quantifying the benefits of using read-across and in silico techniques to fulfill hazard data requirements for chemical categories, *Regul. Toxicol. Pharmacol.* 81 (2016) 250–259. doi:10.1016/j.yrtph.2016.09.004.
- [19] NRC, Toxicity Testing in the Twenty-first Century: A Vision and a Strategy (2007), (2007). <http://dels.nas.edu/Report/Toxicity-Testing-Twenty-first/11970> (accessed March 13, 2017).
- [20] NRC, Toxicity Testing in the 21st Century: A Vision and a Strategy (Report in brief), 2007. http://dels.nas.edu/resources/static-assets/materials-based-on-reports/reports-in-brief/Toxicity_Testing_final.pdf (accessed December 20, 2016).
- [21] M.E. Andersen, D. Krewski, Toxicity Testing in the 21st Century: Bringing the Vision to Life, *Toxicol. Sci.* 107 (2008) 324–330. doi:10.1093/toxsci/kfn255.
- [22] D. Krewski, D. Acosta, M. Andersen, H. Anderson, J.C. Bailar, K. Boekelheide, R. Brent, G. Charnley, V.G. Cheung, S. Green, K.T. Kelsey, N.I. Kerkvliet, A.A. Li, L. McCray, O. Meyer, R.D. Patterson, W. Pennie, R.A. Scala, G.M. Solomon, M. Stephens, J. Yager, L. Zeise, Staff of Committee on Toxicity Test, Toxicity Testing in the 21st Century: A Vision and a Strategy, *J. Toxicol. Environ. Heal. Part B.* 13 (2010) 51–138. doi:10.1080/10937404.2010.483176.
- [23] AOP-42, Inhibition of Thyroperoxidase and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/42> (accessed March 13, 2017).
- [24] AOP-8, Upregulation of Thyroid Hormone Catabolism via Activation of Hepatic Nuclear Receptors, and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/8> (accessed March 13, 2017).
- [25] AOPs, AOPs in AOP-Wiki as of March 2017, (2017). <https://aopwiki.org/aops> (accessed March 13, 2017).
- [26] AOP-Wiki, The AOP-Wiki homepage, (2017). <https://aopwiki.org/> (accessed March 13, 2017).
- [27] OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, (2007). doi:10.1787/9789264085442-en.

PART II - Background

2.1 The Endocrine System and Endocrine Disrupting Chemicals

2.1.1 The Endocrine System

The endocrine system is large and complex and serves multiple essential functions in the body such as regulation of body temperature, blood glucose levels, reproductive function and fetal development [1]. Briefly, the endocrine system ensures optimal communication between cells, tissues and organs of the body through hormone signaling to the responsive tissues. Hormones are synthesized in a number of tissues and organs, a few examples being the thyroid gland, ovaries, testes, hypothalamus, pituitary gland, adrenal glands, adipose tissue and pancreas (Figure 1) [1]. The hormones are released to the bloodstream and transported, often by plasma proteins, to their target tissue(s). Here a hormone can act directly on membrane receptors that transduce signals into the cell or it can enter the cell either by passive diffusion or active transport by membrane proteins [1]. In the cell, the hormone binds and activates its cognate hormone receptor, resulting in downstream effects such as production of proteins that facilitate biological responses [2]. The hormone-receptor interaction pathway is the best-characterized hormone signaling pathway but other modes of action of hormones also exist [3–6].

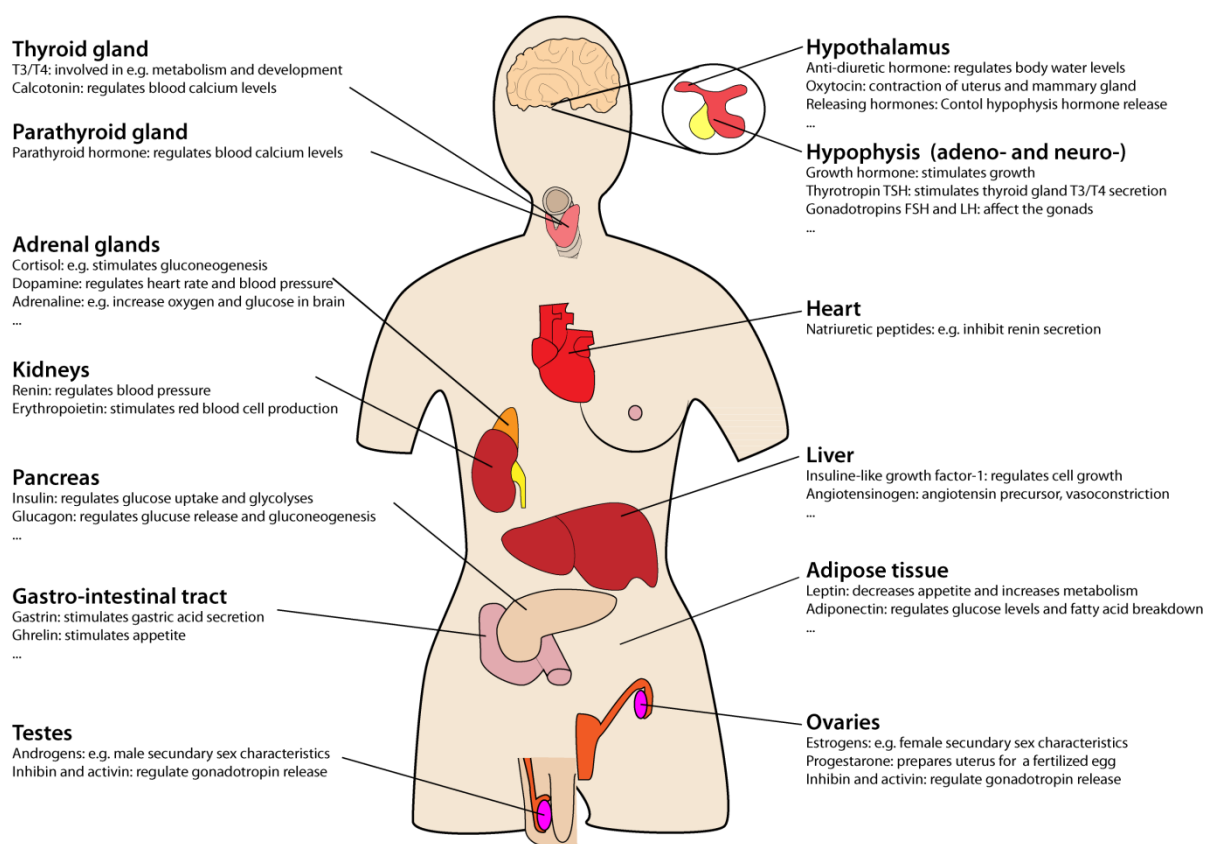


Figure 1. A basic and non-comprehensive overview of the complex endocrine system with examples of hormones and their physiological functions. FSH, follicle stimulating hormone; LH, luteinizing hormone; T4, thyroxine; T3, triiodothyronine; TSH, thyroid stimulating hormone.

The plasma levels of hormones are generally kept within strict, but very individual, patterns by for example negative feedback loops [1,2,6]. With negative feedback loops the hypothalamus, pituitary and in some cases the hormone-producing tissues sense the plasma concentration of the hormone, and in case of a low hormone plasma level synthesis and secretion of the hormone is upregulated and vice versa with a high hormone plasma level. Hormones are metabolized and inactivated by enzymes in the target tissues and/or the liver, and are either reused or excreted via the bile or urine. The expression of the phase I and II liver metabolizing enzymes and the membrane transport proteins is regulated by nuclear receptors (NRs) such as the Pregnane X Receptor (PXR), the Aryl hydrocarbon Receptor (AhR) and Constitutive Androstane Receptor (CAR) [7,8].

2.1.1.1 Thyroid Hormones and Neurodevelopment

Thyroid hormones (THs) are involved in multiple biological processes from early fetal development and throughout adulthood [6,9–11]. In early gestation, the fetus depends on maternally-derived THs. The fetal thyroid gland develops from the third week of gestation, and at approximately gestational week 12 in humans and gestation day 17.5-18 in rats, the fetal thyroid gland starts to synthesize THs from maternally-derived iodine [2,12]. However, maternal THs continue to contribute significantly to fetal TH levels throughout gestation in both humans and rats [10,13]. Consequently, the maternal thyroid gland has to increase its TH production during pregnancy to meet the needs of both fetus and mother [2].

THs are synthesized in the follicles of the thyroid gland located on the anterior trachea (Figure 2a). Serum iodide (I^-) is transported into the thyrocytes by the Na^+/I^- symporter (NIS) in the basal membrane and is further moved across the apical membrane by the anion transporter Pendrin to enter the colloid of the thyroid follicle [14,15]. Here I^- is oxidized to hypoiodite (IO^-) in the presence of dual-oxidase generated hydrogen peroxide (H_2O_2) by the multifunction, heme-containing enzyme thyroperoxidase (TPO) located in the apical thyrocyte membrane [14,16,17]. TPO further catalyzes the iodination of the tyrosyl residues on thyroglobulin (Tg), a glycoprotein secreted by the thyrocytes, to form monoiodotyrosine (MIT) and diiodotyrosine (DIT) [14,16,17]. The conjugation, again catalyzed by TPO, of DITs and MITs on Tg, leads to the formation of three THs: thyroxine (DIT + DIT, T_4), triiodothyronine (MIT + DIT, T_3) or reverse triiodothyronine (DIT + MIT, rT_3), which is biologically inactive [18].

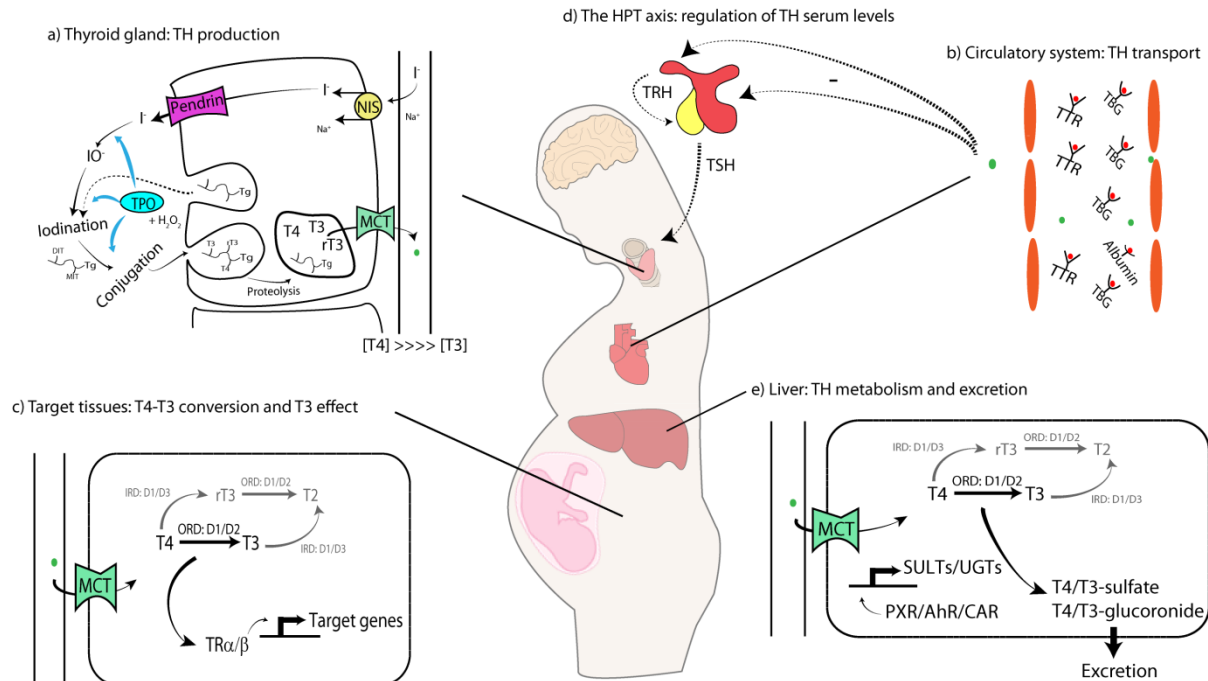


Figure 2. Overview of mechanisms in the thyroid system. See text for explanations and abbreviations.

After being transported across the cell, the THs are released from Tg and secreted into the blood, where the hydrophobic THs are bound to three principal serum TH-binding proteins, thyroxine binding globulin (TBG), transthyretin (TTR) and albumin [19] (Figure 2b). TBG is the main TH plasma transport protein in humans, whereas in animals TTR is the most important transporter protein for THs [2]. TTR also plays a role in the transport of THs over the placenta and the blood-brain-barrier in humans [20,21]. When reaching the target tissue, free serum THs enter the cells by active transporters such as monocarboxylate transporter-8 (MCT-8) and organic anion transporter protein 1c1 (OATP1c1) [10] (Figure 2c and 2e). T4 is the most abundant TH in the blood and is generally converted to the more potent T3 in the liver or locally in the target tissue by outer-ring deiodinase activity (ORD, deiodinase type 1 and 2) [2,10,22]. The effects of T3 is primarily exerted through the two cognate thyroid hormone receptors (TR), TR α and TR β , which bind to thyroid hormone response elements (TREs) to modulate downstream gene transcription resulting in different outcomes depending on the target cell and tissue [10]. Besides regulating TR transcriptional activity, THs can also mediate non-genomic pathways, such as membrane signaling pathways, resulting in rapid (seconds to minutes) onset effects [6].

The TH serum level is normally kept within a narrow range by the hypothalamus-pituitary-thyroid (HPT) axis, a multi-loop negative feedback system that ensures an appropriate balance between synthesis and degradation of THs [2,6] (Figure 2d). In response to low levels of THs in the blood, the pituitary upregulates the secretion of thyroid stimulating hormone (TSH), either as a direct response

or via thyroid releasing hormone (TRH) from the hypothalamus [6]. TSH binds to TSH receptors on the thyrocytes leading to a stimulation of TH synthesis and release [2]. On the other hand, when the TH blood level is high TSH secretion is downregulated resulting in decreased TH synthesis and release. Besides the control of TH levels by the HPT axis, TH levels can also be affected by TH catabolism. THs are primarily metabolized and inactivated in the liver by the phase II enzymes, sulfotransferases (SULTs) and UDP-glucuronosyltransferases (UGTs) [8,23–25], and by inner ring deiodinase activity (IRD, deiodinase type 1 and 3) in both the liver and other tissues [10] (Figure 2e). The expression of SULT and UGT isoenzymes is regulated by the xenobiotic NRs PXR, AhR, and CAR [7,23,26]. The modified and biologically inactive THs are eliminated via the bile or urine.

In adulthood, THs are involved in blood glucose regulation, heart function and basal metabolic rate as well as many other biological processes [27,28]. Dysregulated TH levels can give reversible clinical symptoms of hypo- or hyperthyroidism [28] and are associated with pathological processes involved in adverse outcomes such as cancer, obesity and type II diabetes mellitus [29,30]. In the developing fetus and neonate, THs are involved in various developmental processes [28] and are essential in normal neurodevelopment [2,31]. Both *in vitro* and animal studies have shown the importance of THs in processes such as neuron differentiation, proliferation and migration, dendritic branching and synaptogenesis as well as myelination [10,32,33]. Studies have shown that even a moderate and transient decrease in maternal TH levels during pregnancy is associated with permanent adverse neurological changes in the offspring [2,28]. These changes include reduced IQ and altered cognition, socialization and motor function in children [34–39], and altered cognitive behavior and motor function as well as hearing loss in animals [13,40–42]. Alterations in maternal TH levels during pregnancy, for example due to iodine deficiency or untreated thyroid disorders, have also been associated with an increased risk of cretinism, autism spectrum disorders (ASD) and attention-deficit/hyperactivity disorder (ADHD) in children [9,43–45].

2.1.2 Endocrine Disrupting Chemicals

An endocrine disrupting chemical (EDC) is, as defined by the World Health Organization (WHO) in the International Programme on Chemical Safety (IPCS) report from 2002 [46]:

‘an exogenous substance or mixture that alters function(s) of the endocrine system and consequently causes adverse health effects in an intact organism, or its progeny, or (sub)populations’.

This definition is widely accepted as it is applicable to both human health and ecotoxicological hazard and risk assessment; however it is also relatively open for interpretation. Other definitions of EDCs with focus on the mode-of-actions of EDCs have been suggested [47], for example the EDC definition by Kavlock and others [48]: *‘an exogenous agent that interferes with the production,*

release, transport, metabolism, binding, action or elimination of natural hormones in the body responsible for the maintenance of homeostasis and the regulation of developmental processes’.

Depending on multiple factors such as the timing and length of exposure as well as dose and concurrent exposure to other EDCs, an EDC can modulate the endocrine system and potentially result in adverse effects [1,2,49]. In general, low and transient EDC exposure during adulthood can be compensated for and will often give undetectable or only temporary, reversible effects. Exposure to EDCs during fetal and neonatal development can result in serious and permanent later life effects such as learning disabilities and reduced fertility [1,50]. Because of the complexity of the endocrine system (Figure 1), the cross-talks between the different mechanisms [51,52] and the tempo-spatial aspects, it is difficult to predict if and how endocrine system modulations by EDCs will result in effects at the epi-molecular levels [1]. This is further complicated by interspecies differences in the endocrine effects, which is why extrapolation between results from *in vitro*, *in vivo* and clinical EDC studies should be made with precautions [1].

Multiple programs are screening chemicals for endocrine disrupting properties [32,53,54], and such programs have originally mainly focused on estrogen and androgen receptor interaction. The screening batteries have gradually been extended to cover other endocrine systems such as the thyroid system as well as other mechanisms within the endocrine systems for example the production and degradation of hormones [8,55–59]. The larger the EDC screening battery gets, the better the identification of potential EDCs becomes. Conceptually, one should keep in mind that a chemical can never be said to be without any endocrine modulating potential based on such screenings. Instead, the screenings can help identifying and prioritizing chemicals for further testing/evaluation and aid in the design of higher-tier toxicity testing protocols. They may also provide useful information in combination with AOP(s) to Integrated Approaches and Testing Assessments (IATA) in weight-of-evidence (WoE) assessments as well as give useful information in the substitution to safer alternatives (see chapter 2.3).

2.1.2.1 Thyroid Disrupting Chemicals and Developmental Neurotoxicity

Neurodevelopmental disabilities including ADHD, ASD and IQ deficits are common and their prevalence’s seem to be increasing [60,61]. The causes of neurodevelopmental disabilities are not fully understood, but genetics and environmental factors such as exposure to man-made chemicals are involved [60,61]. Chemicals that interfere with one or more mechanisms in the thyroid system (Figure 2), i.e. thyroid disrupting chemicals (TDCs), can lead to altered TH levels [28]. Studies indicate that the majority of TDCs act by modulating the TH levels rather than direct interaction with the TRs in the target tissues [8]. Exposure to TDCs during pregnancy may lead to decreased maternal TH

levels potentially resulting in developmental neurotoxicity (DNT) and other adverse effects in the offspring [2,7,8,62–65]. Chemical interference with other endocrine and non-endocrine mechanisms may also result in DNT [66,67]. EDCs, and especially TDCs, with DNT potential have been demonstrated to contribute to neurodevelopmental disabilities [60,61,68]. The neurodevelopmental disabilities have multiple implications including reduced life quality and academic achievement, as well as disturbed behavior. These implications have profound economic consequences for societies [60,61], for example is EDC-related DNT estimated to cost Europe more than 150 billion euros per year [68].

Because of the severity of the adverse effects and the economic consequences that can be expected from chemical disruption of thyroid homeostasis there is an urgent need to develop a strategy for the identification and testing of TDCs [8]. This has initiated a large international collaboration, which aims at developing and using new *in vitro* assays for DNT, including *in vitro* assays for thyroid-related mechanisms such as TPO, NIS and deiodinase interaction [66,69]. Such assays can be used for screening the many thousands of chemicals in commerce for which there is none or only limited data on their potential to be TDCs and/or cause DNT. These screening data can be used to either prioritize chemicals for further DNT testing or for inclusion in WoEs of IATAs, e.g. together with relevant AOP(s) and other data, in chemical-specific assessments (see section 2.3.4).

References

- [1] Å. Bergman, J.J. Heindal, S. Jobling, K.A. Kidd, R.T. Zoeller, State of the science of endocrine disrupting chemicals 2012, World Health Organization and United Nations Environment Programme, 2013.
http://apps.who.int/iris/bitstream/10665/78102/1/WHO_HSE_PHE_IHE_2013.1_eng.pdf (accessed March 13, 2017).
- [2] WHO, Endocrine disruptors and child health: Possible developmental early effects of endocrine disruptors on child health, (2012).
http://apps.who.int/iris/bitstream/10665/75342/1/9789241503761_eng.pdf (accessed March 13, 2017).
- [3] N. Heldring, A. Pike, S. Andersson, J. Matthews, G. Cheng, J. Hartman, M. Tujague, A. Strom, E. Treuter, M. Warner, J.-Å. Gustafsson, Estrogen Receptors: How Do They Signal and What Are Their Targets, *Physiol. Rev.* 87 (2007) 905–931. doi:10.1152/physrev.00026.2006.
- [4] S. Nilsson, S. Mäkelä, E. Treuter, M. Tujague, J. Thomsen, G. Andersson, E. Enmark, K. Pettersson, M. Warner, J.-Å. Gustafsson, Mechanisms of Estrogen Action, *Physiol. Rev.* 81 (2001) 1535–1565.
- [5] E.R. Prossnitz, M. Barton, The G-protein-coupled estrogen receptor GPER in health and disease, *Nat. Rev. Endocrinol.* 7 (2011) 715–726. doi:10.1038/nrendo.2011.122.
- [6] R.T. Zoeller, S.W. Tan, R.W. Tyl, General Background on the Hypothalamic-Pituitary-Thyroid (HPT) Axis, *Crit. Rev. Toxicol.* 37 (2007) 11–53. doi:10.1080/10408440601123446.
- [7] AOP-8, Upregulation of Thyroid Hormone Catabolism via Activation of Hepatic Nuclear Receptors, and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017).
<https://aopwiki.org/aops/8> (accessed March 13, 2017).
- [8] A.J. Murk, E. Rijntjes, B.J. Blaauboer, R. Clewell, K.M. Crofton, M.M.L. Dingemans, J. David Furlow, R. Kavlock, J. Köhrle, R. Opitz, T. Traas, T.J. Visser, M. Xia, A.C. Gutleb, Mechanism-based testing strategy using in vitro approaches for identification of thyroid hormone disrupting chemicals, *Toxicol. Vitro.* 27 (2013) 1320–1346. doi:10.1016/j.tiv.2013.02.012.
- [9] Z.-P. Chen, B.S. Hetzel, Cretinism revisited, *Best Pract. Res. Clin. Endocrinol. Metab.* 24 (2010) 39–50. doi:10.1016/j.beem.2009.08.014.
- [10] G.R. Williams, Neurodevelopmental and Neurophysiological Actions of Thyroid Hormone, *J. Neuroendocrinol.* 20 (2008) 784–794. doi:10.1111/j.1365-2826.2008.01733.x.
- [11] P.M. Yen, Physiological and molecular basis of thyroid hormone action., *Physiol. Rev.* 81 (2001) 1097–1142. <http://www.ncbi.nlm.nih.gov/pubmed/11427693>.
- [12] J. Kratzsch, F. Pulzer, Thyroid gland development and defects, *Best Pract. Res. Clin. Endocrinol. Metab.* 22 (2008) 57–75. doi:10.1016/j.beem.2007.08.006.
- [13] K.L. Howdeshell, A Model of the Development of the Brain as a Construct of the Thyroid System, *Environ. Health Perspect.* 110 (2002) 337–348. doi:10.1289/ehp.02110s3337.
- [14] N. Carrasco, Iodide transport in the thyroid gland, *Biochim. Biophys. Acta - Rev. Biomembr.* 1154 (1993) 65–82. doi:10.1016/0304-4157(93)90017-I.
- [15] L. Twyffels, C. Massart, P.E. Golstein, E. Raspe, J. Van Sande, J.E. Dumont, R. Beauwens, V. Kruys, Pendrin: the Thyrocyte Apical Membrane Iodide Transporter?, *Cell. Physiol. Biochem.* 28 (2011) 491–496. doi:10.1159/000335110.
- [16] R.S. Fortunato, E.C. Lima de Souza, R.A. Hassani, M. Boufraquech, U. Weyemi, M. Talbot, O. Lagente-Chevallier, D.P. de Carvalho, J.-M. Bidart, M. Schlumberger, C. Dupuy, Functional

- Consequences of Dual Oxidase-Thyroperoxidase Interaction at the Plasma Membrane, *J. Clin. Endocrinol. Metab.* 95 (2010) 5403–5411. doi:10.1210/jc.2010-1085.
- [17] J. Ruf, P. Carayon, Structural and functional aspects of thyroid peroxidase, *Arch. Biochem. Biophys.* 445 (2006) 269–277. doi:10.1016/j.abb.2005.06.023.
- [18] A. Taurog, M.L. Dorris, D.R. Doerge, Mechanism of Simultaneous Iodination and Coupling Catalyzed by Thyroid Peroxidase, *Arch. Biochem. Biophys.* 330 (1996) 24–32. doi:10.1006/abbi.1996.0222.
- [19] G.C. Schussler, The Thyroxine-Binding Proteins, *Thyroid*. 10 (2000) 141–149. doi:10.1089/thy.2000.10.141.
- [20] R.H. Mortimer, K.A. Landers, B. Balakrishnan, H. Li, M.D. Mitchell, J. Patel, K. Richard, Secretion and transfer of the thyroid hormone binding protein transthyretin by human placenta, *Placenta*. 33 (2012) 252–256. doi:10.1016/j.placenta.2012.01.006.
- [21] S.J. Richardson, R.C. Wijayagunaratne, D.G. D’Souza, V.M. Darras, S.L.J. Van Herck, Transport of thyroid hormones via the choroid plexus into the brain: the roles of transthyretin and thyroid hormone transmembrane transporters, *Front. Neurosci.* 9 (2015) 1–8. doi:10.3389/fnins.2015.00066.
- [22] D.L. St. Germain, V.A. Galton, The Deiodinase Family of Selenoproteins, *Thyroid*. 7 (1997) 655–668. doi:10.1089/thy.1997.7.655.
- [23] K.M. Crofton, Thyroid disrupting chemicals: mechanisms and mixtures, *Int. J. Androl.* 31 (2008) 209–223. doi:10.1111/j.1365-2605.2007.00857.x.
- [24] M.H.A. Kester, E. Kaptein, T.J. Roest, C.H. van Dijk, D. Tibboel, W. Meinel, H. Glatt, M.W.H. Coughtrie, T.J. Visser, Characterization of Human Iodothyronine Sulfotransferases 1, *J. Clin. Endocrinol. Metab.* 84 (1999) 1357–1364. doi:10.1210/jcem.84.4.5590.
- [25] M.H.A. Kester, E. Kaptein, T.J. Roest, C.H. van Dijk, D. Tibboel, W. Meinel, H. Glatt, M.W.H. Coughtrie, T.J. Visser, Characterization of rat iodothyronine sulfotransferases, *Am. J. Physiol. - Endocrinol. Metab.* 285 (2003) E592–E598. doi:10.1152/ajpendo.00046.2003.
- [26] A.H. Tolson, H. Wang, Regulation of drug-metabolizing enzymes by xenobiotic receptors: PXR and CAR, *Adv. Drug Deliv. Rev.* 62 (2010) 1238–1249. doi:10.1016/j.addr.2010.08.006.
- [27] B. Biondi, E.A. Palmieri, G. Lombardi, S. Fazio, Effects of Thyroid Hormone on Cardiac Function - The Relative Importance of Heart Rate, Loading Conditions, and Myocardial Contractility in the Regulation of Cardiac Performance in Human Hyperthyroidism, *J. Clin. Endocrinol. Metab.* 87 (2002) 968–974. doi:10.1210/jcem.87.3.8302.
- [28] M.D. Miller, K.M. Crofton, D.C. Rice, R.T. Zoeller, Thyroid-Disrupting Chemicals: Interpreting Upstream Biomarkers of Adverse Outcomes, *Environ. Health Perspect.* 117 (2009) 1033–1041. doi:10.1289/ehp.0800247.
- [29] E.N. Pearce, Thyroid hormone and obesity, *Curr. Opin. Endocrinol. Diabetes Obes.* 19 (2012) 408–413. doi:10.1097/MED.0b013e328355cd6c.
- [30] C. Wang, The Relationship between Type 2 Diabetes Mellitus and Related Thyroid Diseases, *J. Diabetes Res.* 2013 (2013) 1–9. doi:10.1155/2013/390534.
- [31] R.T. Zoeller, K.M. Crofton, Mode of Action: Developmental Thyroid Hormone Insufficiency—Neurological Abnormalities Resulting From Exposure to Propylthiouracil, *Crit. Rev. Toxicol.* 35 (2005) 771–781. doi:10.1080/10408440591007313.
- [32] E. Ausó, R. Lavado-Autric, E. Cuevas, F.E. del Rey, G. Morreale de Escobar, P. Berbel, A Moderate and Transient Deficiency of Maternal Thyroid Function at the Beginning of Fetal

- Neocortico genesis Alters Neuronal Migration, *Endocrinology*. 145 (2004) 4037–4047. doi:10.1210/en.2004-0274.
- [33] E. Cuevas, E. Ausó, M. Telefont, G.M. de Escobar, C. Sotelo, P. Berbel, Transient maternal hypothyroxinemia at onset of corticogenesis alters tangential migration of medial ganglionic eminence-derived neurons, *Eur. J. Neurosci*. 22 (2005) 541–551. doi:10.1111/j.1460-9568.2005.04243.x.
- [34] P. Berbel, J.L. Mestre, A. Santamaría, I. Palazón, A. Franco, M. Graells, A. González-Torga, G.M. de Escobar, Delayed Neurobehavioral Development in Children Born to Pregnant Women with Mild Hypothyroxinemia During the First Month of Gestation: The Importance of Early Iodine Supplementation, *Thyroid*. 19 (2009) 511–519. doi:10.1089/thy.2008.0341.
- [35] J.E. Haddow, G.E. Palomaki, W.C. Allan, J.R. Williams, G.J. Knight, J. Gagnon, C.E. O’Heir, M.L. Mitchell, R.J. Hermos, S.E. Waisbren, J.D. Faix, R.Z. Klein, Maternal Thyroid Deficiency during Pregnancy and Subsequent Neuropsychological Development of the Child, *N. Engl. J. Med*. 341 (1999) 549–555. doi:10.1056/NEJM199908193410801.
- [36] L. Kooistra, S. Crawford, A.L. van Baar, E.P. Brouwers, V.J. Pop, Neonatal Effects of Maternal Hypothyroxinemia During Early Pregnancy, *Pediatrics*. 117 (2006) 161–167. doi:10.1542/peds.2005-0227.
- [37] Y. Li, Z. Shan, W. Teng, X. Yu, Y. Li, C. Fan, X. Teng, R. Guo, H. Wang, J. Li, Y. Chen, W. Wang, M. Chawinga, L. Zhang, L. Yang, Y. Zhao, T. Hua, Abnormalities of maternal thyroid function during pregnancy affect neuropsychological development of their children at 25-30 months, *Clin. Endocrinol*. 72 (2010) 825–829. doi:10.1111/j.1365-2265.2009.03743.x.
- [38] G. Morreale de Escobar, M. Jesús Obregón, F. Escobar del Rey, Is Neuropsychological Development Related to Maternal Hypothyroidism or to Maternal Hypothyroxinemia? 1, *J. Clin. Endocrinol. Metab*. 85 (2000) 3975–3987. doi:10.1210/jcem.85.11.6961.
- [39] V.J. Pop, J.L. Kuijpers, A.L. van Baar, G. Verkerk, M.M. van Son, J.J. de Vijlder, T. Vulsma, W.M. Wiersinga, H.A. Drexhage, H.L. Vader, Low maternal free thyroxine concentrations during early pregnancy are associated with impaired psychomotor development in infancy, *Clin. Endocrinol*. 50 (1999) 149–155. doi:10.1046/j.1365-2265.1999.00639.x.
- [40] K.M. Crofton, Developmental Disruption of Thyroid Hormone: Correlations with Hearing Dysfunction in Rats, *Risk Anal*. 24 (2004) 1665–1671. doi:10.1111/j.0272-4332.2004.00557.x.
- [41] E.S. Goldey, L.S. Kehn, G.L. Rehnberg, K.M. Crofton, Effects of Developmental Hypothyroidism on Auditory and Motor Function in the Rat, *Toxicol. Appl. Pharmacol*. 135 (1995) 67–76. doi:10.1006/taap.1995.1209.
- [42] R.T. Zoeller, J. Rovet, Timing of Thyroid Hormone Action in the Developing Brain: Clinical Observations and Experimental Findings, *J. Neuroendocrinol*. 16 (2004) 809–818. doi:10.1111/j.1365-2826.2004.01243.x.
- [43] S. Andersen, P. Laurberg, C. Wu, J. Olsen, Attention deficit hyperactivity disorder and autism spectrum disorder in children born to mothers with thyroid dysfunction: a Danish nationwide cohort study, *BJOG* 121 (2014) 1365–1374. doi:10.1111/1471-0528.12681.
- [44] S. Hoshiko, J.K. Grether, G.C. Windham, D. Smith, K. Fessel, Are thyroid hormone concentrations at birth associated with subsequent autism diagnosis?, *Autism Res*. 4 (2011) 456–463. doi:10.1002/aur.219.
- [45] T. Modesto, H. Tiemeier, R.P. Peeters, V.W. V. Jaddoe, A. Hofman, F.C. Verhulst, A. Ghassabian, Maternal Mild Thyroid Hormone Insufficiency in Early Pregnancy and Attention-Deficit/Hyperactivity Disorder Symptoms in Children, *JAMA Pediatr*. 169 (2015) 838–845. doi:10.1001/jamapediatrics.2015.0498.

- [46] WHO/IPCS, Global assessment of the state-of-the-science of endocrine disruptors, World Heal. Organ. (2002).
http://www.who.int/ipcs/publications/new_issues/endocrine_disruptors/en/ (accessed March 13, 2017).
- [47] R.T. Zoeller, T.R. Brown, L.L. Doan, A.C. Gore, N.E. Skakkebaek, A.M. Soto, T.J. Woodruff, F.S. Vom Saal, Endocrine-Disrupting Chemicals and Public Health Protection: A Statement of Principles from The Endocrine Society, *Endocrinology*. 153 (2012) 4097–4110. doi:10.1210/en.2012-1422.
- [48] R.J. Kavlock, G.P. Daston, C. DeRosa, P. Fenner-Crisp, L.E. Gray, S. Kaattari, G. Lucier, M. Luster, M.J. Mac, C. Maczka, R. Miller, J. Moore, R. Rolland, G. Scott, D.M. Sheehan, T. Sinks, H.A. Tilson, Research needs for the risk assessment of health and environmental effects of endocrine disruptors: a report of the U.S. EPA-sponsored workshop, *Environ. Health Perspect.* 104 (1996) 715–740.
- [49] E. Diamanti-Kandarakis, J.-P. Bourguignon, L.C. Giudice, R. Hauser, G.S. Prins, A.M. Soto, R.T. Zoeller, A.C. Gore, Endocrine-Disrupting Chemicals: An Endocrine Society Scientific Statement, *Endocr. Rev.* 30 (2009) 293–342. doi:10.1210/er.2009-0002.
- [50] P. Grandjean, P.J. Landrigan, Neurobehavioural effects of developmental toxicity, *Lancet Neurol.* 13 (2014) 330–338. doi:10.1016/S1474-4422(13)70278-3.
- [51] P. Duarte-Guterman, L. Navarro-Martín, V.L. Trudeau, Mechanisms of crosstalk between endocrine systems: Regulation of sex steroid hormone synthesis and action by thyroid hormones, *Gen. Comp. Endocrinol.* 203 (2014) 69–85. doi:10.1016/j.ygcen.2014.03.015.
- [52] C.P. Martucci, J. Fishman, P450 enzymes of estrogen metabolism, *Pharmacol. Ther.* 57 (1993) 237–257. doi:10.1016/0163-7258(93)90057-K.
- [53] EDSP, Federal Register: Environmental Protection Agency - Endocrine Disruptor Screening Program (EDSP); Announcing the Availability of the Tier 1 Screening Battery and Related Test Guidelines; Notice, 2009. <https://www.federalregister.gov/documents/2009/10/21/E9-25348/endocrine-disruptor-screening-program-edsp-announcing-the-availability-of-the-tier-1-screening> (accessed January 19, 2017).
- [54] EDSTAC, Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) Final Report, (1998). <https://www.epa.gov/endocrine-disruption/endocrine-disruptor-screening-and-testing-advisory-committee-edstac-final> (accessed March 13, 2017).
- [55] A.L. Karmaus, C.M. Toole, D.L. Filer, K.C. Lewis, M.T. Martin, High-Throughput Screening of Chemical Effects on Steroidogenesis Using H295R Human Adrenocortical Carcinoma Cells, *Toxicol. Sci.* 150 (2016) 323–332. doi:10.1093/toxsci/kfw002.
- [56] R.J. Kavlock, D. Dix, K. Houck, R. Judson, T. Knudsen, D. Reif, M. Martin, Biological Profiling of Endocrine Related Effects of Chemicals in ToxCast, (2009).
https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=203432&keyword=&actType=&TIMSType=+&TIMSSubTypeID=&DEID=&epaNumber=&ntisID=&archiveStatus=Both&ombCat=Any&dateBeginCreated=&dateEndCreated=&dateBeginPublishedPresented=&dateEndPublishedPresented=&dateBeginUpdated=&dateEndUpdated=&dateBeginCompleted=&dateEndCompleted=&personID=12250&role=Any&journalID=&publisherID=&sortBy=title&count=25&CFID=57839251&CFTOKEN=60543589 (accessed March 13, 2017).
- [57] OECD, New scoping document on in vitro and ex vivo assays for the identification of modulators of thyroid hormone signalling, (2014).
[http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2014\)23&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2014)23&doclanguage=en) (accessed March 13, 2017).

- [58] K. Paul Friedman, E.D. Watt, M.W. Hornung, J.M. Hedge, R.S. Judson, K.M. Crofton, K.A. Houck, S.O. Simmons, Tiered High-Throughput Screening Approach to Identify Thyroperoxidase Inhibitors Within the ToxCast Phase I and II Chemical Libraries, *Toxicol. Sci.* 151 (2016) 160–180. doi:10.1093/toxsci/kfw034.
- [59] D.M. Rotroff, D.J. Dix, K.A. Houck, T.B. Knudsen, M.T. Martin, K.W. McLaurin, D.M. Reif, A. V. Singh, M. Xia, R. Huang, R.S. Judson, Using in Vitro High Throughput Screening Assays to Identify Potential Endocrine-Disrupting Chemicals, *Environ. Health Perspect.* 121 (2012) 7–14. doi:10.1289/ehp.1205065.
- [60] P. Grandjean, P.J. Landrigan, Neurobehavioural effects of developmental toxicity, *Lancet Neurol.* 13 (2014) 330–338. doi:10.1016/S1474-4422(13)70278-3.
- [61] P. Grandjean, P. Landrigan, Developmental neurotoxicity of industrial chemicals, *Lancet.* 368 (2006) 2167–2178. doi:10.1016/S0140-6736(06)69665-7.
- [62] AOP-134, Sodium Iodide Symporter (NIS) Inhibition and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/134> (accessed March 13, 2017).
- [63] AOP-152, Interference with thyroid serum binding protein transthyretin and subsequent adverse human neurodevelopmental toxicity, (2017). <https://aopwiki.org/aops/152> (accessed March 13, 2017).
- [64] AOP-42, Inhibition of Thyroperoxidase and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/42> (accessed March 13, 2017).
- [65] AOP-54, Inhibition of Na⁺/I⁻ symporter (NIS) decreases TH synthesis leading to learning and memory deficits in children, (2017). <https://aopwiki.org/aops/54> (accessed March 13, 2017).
- [66] A. Bal-Price, K.M. Crofton, M. Leist, S. Allen, M. Arand, T. Buetler, N. Delrue, R.E. FitzGerald, T. Hartung, T. Heinonen, H. Hogberg, S.H. Bennekou, W. Lichtensteiger, D. Oggier, M. Paparella, M. Axelstad, A. Piersma, E. Rached, B. Schilter, G. Schmuck, L. Stoppini, E. Tongiorgi, M. Tiramani, F. Monnet-Tschudi, M.F. Wilks, T. Ylikomi, E. Fritsche, International STakeholder NETwork (ISTNET): creating a developmental neurotoxicity (DNT) testing road map for regulatory purposes, *Arch. Toxicol.* 89 (2015) 269–287. doi:10.1007/s00204-015-1464-2.
- [67] A. Bal-Price, K.M. Crofton, M. Sachana, T.J. Shafer, M. Behl, A. Forsby, A. Hargreaves, B. Landesmann, P.J. Lein, J. Louisse, F. Monnet-Tschudi, A. Paini, A. Rolaki, A. Schrattenholz, C. Suñol, C. van Thriel, M. Whelan, E. Fritsche, Putative adverse outcome pathways relevant to neurotoxicity, *Crit. Rev. Toxicol.* 45 (2015) 83–91. doi:10.3109/10408444.2014.981331.
- [68] M. Bellanger, B. Demeneix, P. Grandjean, R.T. Zoeller, L. Trasande, Neurobehavioral Deficits, Diseases, and Associated Costs of Exposure to Endocrine-Disrupting Chemicals in the European Union, *J. Clin. Endocrinol. Metab.* 100 (2015) 1256–1266. doi:10.1210/jc.2014-4323.
- [69] OECD/EFSA, OECD/EFSA Workshop on Developmental Neurotoxicity (DNT): the use of non-animal test methods for regulatory purposes, (2016). <https://www.efsa.europa.eu/en/events/event/161018b> (accessed February 23, 2017).

2.2 Quantitative Structure-Activity Relationship Models

A QSAR model is a mathematical model that describes the quantitative relationship between chemical structures and their properties, e.g. a physico-chemical property or a biological activity. QSARs are trained on experimental data for chemicals with known structures using machine learning and statistical methods, and they can be used to predict the activity of chemicals based on their structures (see e.g. [1] and [2] for more in-depth reviews of QSARs). The *quantitative* in QSAR refers to the nature of the descriptors (i.e., independent variables) and the modeling method and not to the modeled endpoint (i.e., response variable), which can be either quantitative/continuous (e.g. IC_{50}) or qualitative/categorical (e.g. active versus inactive) [1]. Closely related to QSAR is the simpler, structure-activity relationship (SAR) method that qualitatively relates a (sub)structure to an activity. In contrast with QSARs that result from statistical analyses of experimental data, SARs are usually based on expert knowledge and are encoded into expert systems [3]. Collectively, SARs and QSARs are referred to as (Q)SARs [1]. (Q)SARs are non-testing approaches and other related non-testing approaches include grouping approaches using e.g. read-across, and expert systems, which can be combinations of SARs, QSARs and databases [1]. Together these non-testing approaches are based on the structural similarity principle, i.e. the hypothesis that structurally similar chemicals exhibit similar behavior (in living organisms), and are used to facilitate the evaluation of properties of chemicals by extending existing information [1].

QSARs date back to the late 1800s, when Hans Horst Meyer and Fritz Baum described the correlation between partition coefficients and tadpole alcohol narcosis [4–6]. The interest in QSARs has increased gradually [2,7] after the pioneering work in the 1960s by Corwin Hansch and colleagues, who made simple QSAR models for inhibition of photosynthesis and activity of auxin, a plant growth substance [8–11]. Since then advances in technology, mathematical methods, and computer power have allowed for efficient development of much more complex and predictive QSAR models. Today QSAR models are widely used in academia, industry and agencies [2].

2.2.1 QSAR Development

The development of QSAR models follows a general workflow starting with 1) dataset collection, curation and preparation, 2) generation and selection of chemical descriptors to be used as independent variables in 3) the model building step, and finally 4) a statistical validation of the model(s) within the defined applicability domain (AD) (Figure 3) [12]. A QSAR model is built using a so-called training set, which consists of chemical structures and related experimental data. The chemical structures are represented by chemical descriptors (see more in 2.3.1.2), which are used as independent variables in the model. The experimental endpoint, which can be either continuous or

categorical, is used as the response variable in the model. QSARs are normally classified as global or local. A local QSAR is trained on a small and congeneric series of chemical structures, whereas a global QSAR is trained on a large and structurally diverse set of chemicals. The term validation is broadly defined as “the process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose” [13]. However, this definition is rather abstract in a QSAR context and therefore a more operational definition of validation has been proposed [14]: “The validation of a (Q)SAR is the process by which the performance and mechanistic interpretation of the model are assessed for a particular purpose”. The performance assessment here refers to the statistical validation of the model [1]. The AD as a general term is defined as “the response and chemical structure space in which the model makes predictions with a given reliability” [15] (see more in section 2.2.2).

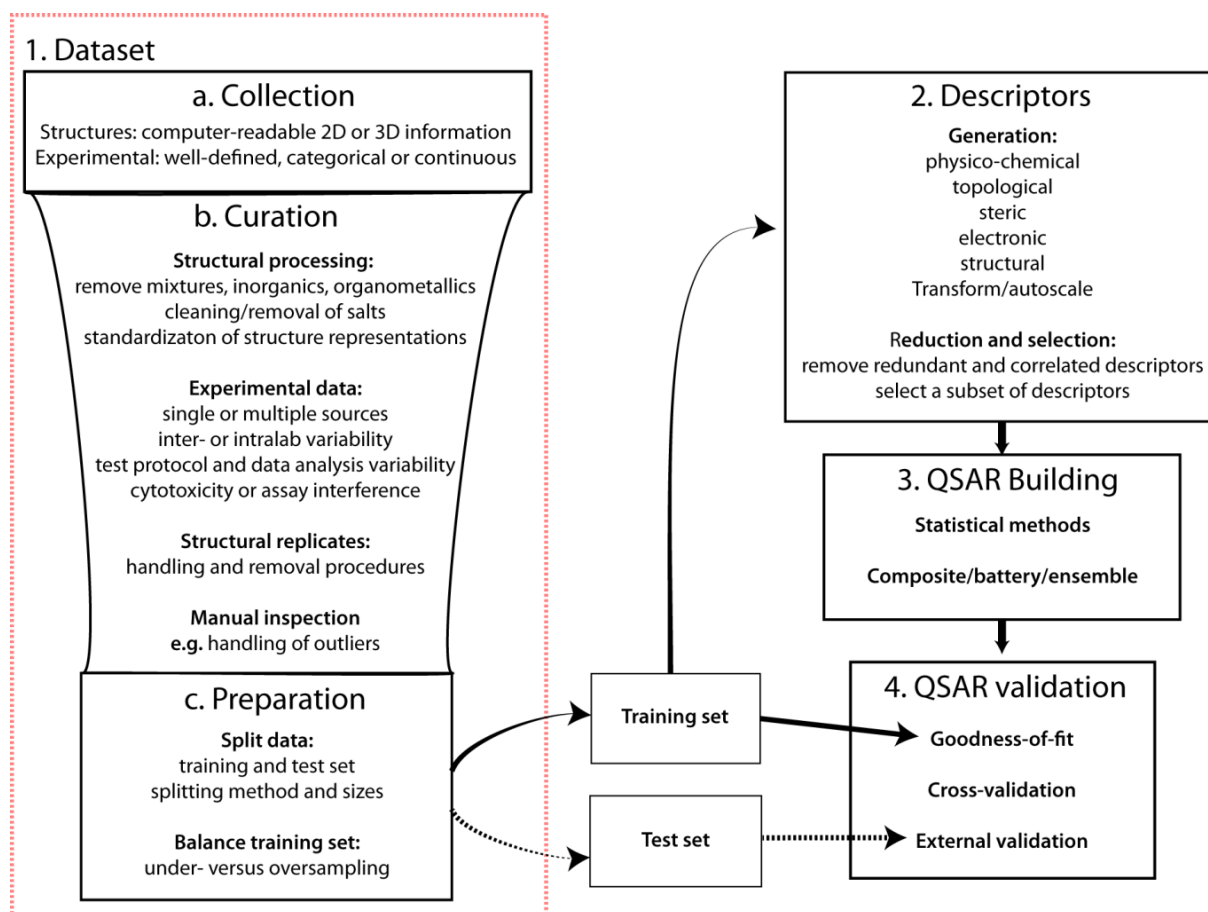


Figure 3: Overview of the basic QSAR development steps. See text for an explanation of the figure.

Many different methods for each of the steps in 1) to 4) have been proposed and used (see e.g. [7,12,16,17]). Here a basic and non-comprehensive workflow is introduced and some of the methods and caveats are briefly discussed (Figure 3).

2.2.1.1 Data Collection, Curation and Preparation

As the quality of the data strongly influences the quality and performance of the built QSAR model [16,18–20], the steps of data collection, curation and preparation are of high importance in QSAR development and should follow some basic principles [2] (Figure 3, 1).

2.2.1.1.1 Data Collection

The first step when developing a QSAR model is to collect a dataset containing structure and experimental endpoint information for a set of chemicals [12,16,21–23]. The chemical structures in the dataset should be represented in a computer-readable 2D or 3D format, three of the most widely used ones being SMILES (simplified molecular input line entry system) [24–26], the connection table format (used in MOL or SDF files) [27] and InCHI (IUPAC International Chemical Identifier)¹. Most QSAR models use 2D structure information but 3D-QSAR models also exists [28–31] and 4- and higher-dimensional approaches have been reported [32]. Preferably the dataset should be collected from a single reliable source and have experimental data for a well-defined endpoint that have been produced in the same laboratory by the same personnel and have followed the same experimental protocol(s) and subsequent data analyses [1] (Figure 3, 1a).

2.2.1.1.2 Data Curation

At this step, the quality of both structure and experimental information in the collected dataset should be thoroughly evaluated as errors in the data can strongly influence the performance of the developed model [2,12,16,21–23,33]. Several studies have shown that structural errors are not uncommon, and therefore identification and correction of such structural errors should be part of a standardized data curation strategy [12,23,33,34]. Often the software systems used for interpreting the structures and/or building the models are limited in the chemical universe they can handle, and most QSAR models are based on organic discrete 2D chemical structures. Inorganic or organometallic compounds and mixtures can generally not be handled by conventional cheminformatics tools and need removal [2]. The remaining structures need to undergo a standardization and normalization procedures to ensure that all structures are described following the same algorithm, i.e. are canonized, in terms of e.g. ring aromatization and neutralization [2,26]. When these steps have been applied the chemical structures are made ‘QSAR-ready’ (Figure 3, 1b).

The quality and reproducibility of the experimental data should be assessed. In general, with regards to the quality and reproducibility of the experimental data the model developer has to rely on the information from the data provider(s). Often a description of the experimental protocol(s) and performance as well as the applied data analyses is available to the model developer. The model

¹ <https://iupac.org/who-we-are/divisions/division-details/inchi/>

developer should become familiar with the nature of the experimental data and its underlying biology and assay technology to assess the degree of uncertainty/artefacts and potential false results. Based on this, measures should be taken to identify unreliable experimental results. For example corresponding experimental data from counterscreen assays, e.g. for luciferase interaction, can be applied to identify non-specific and potential false experimental results. If the data have been collected from multiple sources it can contain additional uncertainties from e.g. interlaboratory variability and/or differences in the test protocols and data analyses. Such uncertainties are likely to introduce extra noise and reduce the performance of the model compared to models built from data from a single source undergoing the same test protocol(s) and data analysis [1,7,35] (Figure 3, 1b).

Next, any replicated 'QSAR-ready' structures in the dataset should be identified and the experimental values of the identified structural replicates should be compared [2]. If the replicates have the same experimental results then only one of the structures should be kept as they will otherwise be given too large influence in the model. For a set of replicates with discrepant experimental results different removal approaches can be used, e.g. removing all replicates or, for a continuous response variable, an average value can be calculated and kept together with one of the structures [2,36] (Figure 3, 1b).

When the replicates have been handled, a final general manual inspection of the dataset should be made as the last step of the data curation and can include checking that previous curation steps have been successful and identification of outliers [2,16] (Figure 3, 1b). Manual inspection is however not practical in case of very large datasets and may be skipped in such cases. Briefly, outliers can be of the 'activity cliff' type or due to errors in structure information or experimental data not taking into consideration in the previous steps [21,37]. There are different approaches about outliers. In principle, all available experimental data are valuable and should be used in the construction of a QSAR model. However, outlier removal if done independent of the model results can be justified in some situations. For example in the development of smaller, local QSAR models based on a dataset of congeneric chemical structures that act by a common mechanism, a correct experimental result may be treated as an outlier if it is known that the chemical acts by a different mechanism than the one for the majority of the training set. Overall, if measures are taken to remove outliers, a good explanation should be provided along with a detailed documentation of the removal procedure, otherwise the outlier removal step can be interpreted as a manipulation of data with the purpose of artificially improving model performance [2,38].

For preparing a prediction set, i.e. a dataset containing only chemical structures that are planned for screening through the QSAR model to generate predictions for their activity, the data collection and curation steps regarding structures also apply.

2.2.1.1.3 Data Preparation

When the data have been properly curated, the next step is to decide whether the curated dataset should be used as a training set or if it should be split into training and test sets [12,39] (Figure 3, 1c). In the last case, different splitting methods can be used such as random splitting or a rational stratified splitting on endpoint activity or descriptor space [17,39–41]. Each splitting method will have its pros and cons with regard to model coverage and performance as well as the interpretation of the external validation estimations [40,42]. In general, rational splitting will result in a test set more similar to the training set and may, if the test set is too similar, give overoptimistic future predictive performance and coverage measures compared to a test set made from random splitting, which better represents the future non-selected prediction sets [2,36,40]. Some things to consider if the dataset is split, besides the splitting method, are the absolute and relative sizes of the training and test sets. Often the size of the test set(s) is between 10% and 30% of the dataset [39], and the remainder of 70% to 90% is used for model training. The absolute size of the test set should be large enough to be used for robust external validation [12,17,43], and similarly the training set should, at least for global QSAR development, have a certain size and diversity to avoid chance correlations and overfitting. Fixed cut-offs for the lower limits of the size of the training and test sets cannot be set [12] as this depends on the nature of the full dataset, the types of chemical descriptors and statistical methods being used, the purpose of the model etc. [39].

Due to the increasing implementation use of high-throughput screening (HTS) assays such as those applied in e.g. the ToxCast and Tox21 programs (see section 2.3.2), it is more and more common to find datasets with a binary response variable that are very imbalanced towards a larger class of inactives [12,44]. Generally, a QSAR model trained on such imbalanced dataset has a tendency to be biased towards making predictions for the majority class (Figure 3, 1c). For the typical imbalanced training set with a bigger inactivity class this will likely result in a model with a high specificity and a low sensitivity (see definitions in Figure 4) upon predictive performance evaluation [45]. To overcome this problem different approaches to balance the training set have been suggested [12,44,46,47] such as undersampling of the bigger class or oversampling of the smaller class [48,49].

2.2.1.2 Descriptor Generation and Selection

To build a QSAR model a set of descriptors encoded within chemical structures of the curated training set first needs to be generated (Figure 3, 2). Chemical descriptors are values that describe

different properties of a molecule. They can be physico-chemical characteristics (e.g. molecular weight and logP), topological (e.g. atom, bond and ring counts), steric (e.g. volume and surface) or electronic (e.g. HOMO and LUMO). A special class of descriptors is structural descriptors or features. Pre-defined sets of structural features, also called structural keys (e.g. MACCS keys [50]) and the Leadscape Structural Feature Hierarchy [51]), can be used by searching for the pre-defined structural keys in the chemical structures of the training set. The presence or absence of each key in a structure is encoded in a bitmap, where each bit represents a 1 if the key is present and a 0 if it is absent. The structural keys can also be used for constructing new, larger structural features [52]. Furthermore, structural features can be molecule-dependent, i.e. so-called fingerprints, rather than pre-defined. Structural descriptors in fingerprints are created using a fingerprinting algorithm (e.g. the Daylight fingerprints [53]) that examines the molecule and generates a set of patterns [54]. Besides being applicable in QSAR modeling, structural features from both structural keys and fingerprints are also used to calculate structural similarity measures such as the Euclidean distance or the Tanimoto/Jaccard coefficient [53,55]. Transformation, i.e. normalization and/or autoscaling, of continuous chemical descriptors and/or the response variable might be necessary at this step as large variabilities in the range and distribution of these can pose a problem for some statistical/machine learning methods [2,7]. Examples of commercial and free software tools for generating chemical descriptors include MOE², DRAGON³, RDKit⁴, PaDEL⁵ and CDK [7,56–59].

The number of generated chemical descriptors for a training set is often huge and many of the descriptors may be correlated or redundant (Figure 3, 2). Examples of redundant descriptors include those only present in a single structure or descriptors with the same or almost same value over all samples in the dataset. Different unsupervised data reduction techniques for removing or minimizing redundant information are available, an example being the principal component analysis (PCA) that creates uncorrelated latent (i.e., hidden/non-observable) variables from the descriptors [54,60,61]. After removing redundant and correlated descriptors, the next step is to select the descriptors that should be included in the model algorithm. Multiple descriptor selection techniques are applied in QSAR development, all with the purpose of finding a combination of descriptors for QSAR modeling of the response variable [62–64]. The selection techniques include supervised methods such as wrapper methods (e.g. genetic algorithms (GAs)), and filter methods (e.g. univariate data analysis) [65]. Each method has its advantages and limitations in terms of e.g. computation time and ease of implementation [62–64,66]. The descriptor reduction and selection

² https://www.chemcomp.com/MOE-Cheminformatics_and_QSAR.htm

³ http://www.taletе.mi.it/products/dragon_description.htm

⁴ <http://www.rdkit.org/>

⁵ <http://www.yapcwsoft.com/dd/padeldescriptor/>

procedures have been used to reduce computation time, improve model predictive performance, and ease interpretability as well as avoid overfitting and reduce chance correlations [62,63,67]. A general recommendation is that the training set chemicals to chemical descriptors ratio of a model should be at least 5:1 in order to minimize the risk of chance correlations and overfitting [2,35,68].

2.2.1.3 Machine Learning and Statistical Methods in QSAR building

Various QSAR modeling methods exist, and new methods are continuously being developed [69] (Figure 3, 3). Depending on whether the response variable is continuous or categorical either regression or classification methods, respectively, should be applied in the QSAR building. In general, classification models tend to be more flexible and successful in prediction [70]. A continuous response variable can be made categorical by using one (i.e. binary) or more cut-offs, which can be set based on different criteria such as model performance or a biological rationale [12,41,71].

The list of classification and regression methods applied in QSAR building is long [69,72]. A few non-exhaustive examples of linear and non-linear classification and regression methods used to build QSAR models are listed in Table 1.

Table 1. Examples on the use of classification and regression methods in QSAR building

Examples	Use cases	References
Classification	Linear discriminant analysis (LDA)	[39]
	<i>k</i> -nearest neighbors (<i>k</i> NN)	[71]
	Naïve bayes (NB) classification	[73]
	Support vector machines (SVM)	[73,74]
	Random forest (RF)	[73,74]
	Partial logistic regression (PLR)	[36]
Regression	Classification and regression trees (CART)	[75]
	Multiple linear regression (MLR)	[39]
	Partial least squares (PLS) regression	[39]
	Artificial neural network (ANN) regression	[66]
	Stepwise regression	[39]

Some of these methods, e.g. RF, SVM and ANN, have been invented and implemented to handle both regression and classification problems [73]. Each method has its advantages and limitations in terms of e.g. computation time/memory, overfitting tendencies, sensitivity to noise and interpretability [69,72,73], and their predictive success depends on the nature of the training set and the types of chemical descriptors. The descriptor selection techniques and QSAR modeling methods are in some cases integrated, e.g. when applying GAs on MLR or SVM [76].

QSARs built using the same training set may produce discrepant predictions for a query chemical due to differences in the applied statistical methods and/or chemical descriptor sets. Therefore, rather

than relying on a single prediction for a given endpoint for the query chemical, increased certainty in the prediction can be achieved by applying a consensus or battery approach (Figure 3, 3). In a consensus or battery approach, the predictions from the individual models are integrated to output one consensus or battery prediction. This approach is used on a large scale e.g. in the Danish (Q)SAR Database (see 2.2.3) [77,78]. In general, by combining multiple predictions to reach final battery/consensus predictions, a better and more correct description of the relationship between the query chemical structure and its predicted activity can be obtained due to the noise or limited coverage of the single model being canceled by the others [79].

2.2.1.4 Methods in Statistical Validation of QSAR Model

After a model has been developed it should be statistically validated for its goodness-of fit, robustness and predictive performance within one or more defined ADs (see more in section 2.2.2). Here some of the most common methods in QSAR validation are briefly presented (Figure 3, 4).

Goodness-of-fit

The goodness-of-fit is a measure of the model's internal performance, i.e. how well the model predicts its own training set. For classification models the goodness-of-fit is sometimes expressed as Cooper statistics [80], including sensitivity, specificity, concordance and balanced accuracy, which are calculated based on the confusion matrix (Figure 4).

Confusion matrix		QSAR Predictions	
		Positive	Negative
Experimental values	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True negatives (TN)

$Sensitivity = \frac{TP}{TP + FN}$, the proportion of experimental actives correctly predicted

$Specificity = \frac{TN}{TN + FP}$, the proportion of the experimental inactives correctly predicted

$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2}$, the average of the sensitivity and specificity

Figure 4: Confusion matrix and Cooper statistics.

External validation

External validation is part of model predictivity assessment and the procedure consists of predicting a test set, i.e. a set of substances not used for training the model. A robust external validation, i.e. made with a test set of sufficient size and structural diversity to be representative of the chemical diversity of the model's training set, is by some scientists considered the 'gold standard' to assess a model's predictive performance (as discussed in [17]). The experimental data of the test set should

preferably be of the same type as the experimental results in the training set, i.e. in the ideal case tested following the same protocol and data analysis at the same laboratory and by the same personnel [35]. When the test set has been run through the model the predictions that fall within the defined AD are compared with the corresponding experimental data and different statistical measures can be calculated, e.g. the Cooper statistics (Figure 4).

A limitation sometimes met by model developers is the absence of a test set. A test set can be acquired by: using part of the curated dataset, i.e. splitting; generating new experimental data; or finding new data in databases or the literature that is similar to the data in the training set. If the entire dataset has been used for model training and new data are not available, external validation is not possible. If the test set is acquired by splitting the experimental dataset into training and test sets, the splitting method is of importance [17]. Robust external validation with a test set from a rational splitting will likely result in more optimistic coverage and predictive performance estimates compared to the estimates from a test set made from random splitting [17]. The external validation results from a test set made with random splitting will generally give more realistic estimates of the model's future screening set performance [40].

Cross-validation

Cross-validation is a common and popular technique used for assessing both model robustness and predictive performance. Cross-validation approaches include for example leave-one-out (LOO), leave-many-out (LMO), randomization, stratified randomization and bootstrapping. LOO is a type of k -fold cross-validation, which is a commonly used cross-validation method for QSAR models. Briefly, in k -fold cross-validation, the training set, S , is split into k subsets S_1, \dots, S_k , where

$$\bigcup_{i=1}^k S_i = S, \bigcap_{i=1}^k S_i = \emptyset$$

In the LOO case, k is equal to the number of entries in the training set. The selected k is often dependent on the training set size, and regularly used k for robust cross-validations includes 2, 5, 10 and 20 [17,36]. Then k cross-validation models, M_i , are built using $S \setminus S_i$, so that all k subsets have been included in all but one of the k cross-validation models. Each cross-validation model, M_i , which should be built without any transfer of information from the full parent model such as selected descriptors, is externally validated with the left-out subset, S_i . The procedure can be made x times in a so-called x times k -fold cross-validation. The statistical results from the k external validations are averaged to give an overall statistics, which is then used as an estimate for the predictive performance of the parent model made on the full-training set, S (Figure 4). The variance in the individual cross-validation model performance measures, expressed as e.g. a standard deviation

(SD), can be used for estimating the robustness of the parent model. Large variability, i.e. high SDs, in the cross-validation model performance estimates indicate a parent model being easily affected by changes in the constitution of the training set.

Some scientists criticize the use of cross-validation to assess model predictivity [81,82] as results from cross-validations have in some cases reported optimistic and misleading estimates. Such optimistic results are likely derived from cross-validations where either k has been too large, e.g. LOO on large training sets, or information from the full training set model has been transferred to the cross-validation models [65]. They may also be due to conservative measures derived from an external validation with uncritical use of a test set with experimental results that are not similar enough to the training set experimental results. A large systematic study that compared robust cross-validation, i.e. no reuse of information and appropriate sizes of k , with robust external validation has shown that robust cross-validation generally underestimated model predictivity [17].

To summarize on the topic of statistical validation of QSAR models, a combination of robust external and cross-validation is likely the optimal, although not always a practical, choice when assessing the robustness and predictive performance of a model [39]. If part of the dataset has been used as a test set for robust external validation, this can have an effect on the developed model, which can suffer on both coverage and predictive performance of future screening sets [17] due to the resulting lower number of chemicals available for model training. To circumvent this in practice, the test set can be added to the training set and used for building a bigger model. In this case it is important to remember that the external validation results from the first model do not apply on the new bigger model. However, by comparing the results from the external validation with cross-validation of the first model, an indication can be obtained of whether the cross-validation procedure outputs realistic results or if it is either overoptimistic or conservative in its nature. This information can be taken into consideration when assessing the cross-validation results of the bigger models. A comparison of corresponding measures from the goodness-of-fit and the external- and/or cross-validation can be made. If the statistical measures from the goodness-of-fit test are significantly larger than those from the external- and/or cross-validation, this indicates that the model has been overfitted to its training set and thus lost some of its ability to generalize. Overfitting may be due to inclusion of too many descriptors in the model, or it can be related to the model building method and its parameters [83].

2.2.1.5 QSAR Development using Leadscape Predictive Data Miner

In this PhD project, the commercial QSAR modeling software Leadscape Predictive Data Miner (LPDM), a component of LeadScope®Enterprise Server⁶, was used for 2D QSAR development. The data collection, curation and preparation steps were made prior to the import of datasets into LPDM using programs such as Microsoft Excel and OASIS Database Manager [84]. OASIS Database Manager is a software platform that can store chemical structures as well as process and manage chemical information [84]. Here is a brief and more theoretical description of LPDMs QSAR development methods. More detailed descriptions on the practical use of LPDM in the PhD project are given in the respective project chapters in Part III.

During the import of a dataset into LPDM, a set of nine molecular descriptors are automatically calculated for each structure: ALogP, Hydrogen Bond Acceptors and Donors, Lipinski Score, Molecular Weight, Parent Atom Number, Parent Molecular Weight, Polar Surface Area, and Rotatable Bonds. Additionally, a systematic substructure analysis is performed on each structure using a hierarchy of approximately 27,000 pre-defined 2D structural keys [51,52,85,86].

When a training set has been successfully imported, model development can be started and consists of three main steps. In the first and optional step more descriptors can be added to the initial descriptor set prepared in the importing step, i.e. the pre-defined structural features and the calculated molecular descriptors. The new descriptors can come from the generation of predictive scaffolds from the current dataset, addition of previously generated dataset scaffolds or by importing descriptors from an external source. The scaffolds are created by assembling LPDM pre-defined structural keys into larger substructures that are commonly occurring within a group of training set structures or that discriminate for the response variable [52,86]. In LPDM, the descriptor selection is divided into two phases: 1) a pre-selection of descriptors before model building, and 2) an iterative descriptor reduction during model building to optimize the number of descriptors and factors (i.e., latent variables) in the model [85].

The second step of LPDM model development includes the phase 1) pre-selection of descriptors from the calculated molecular descriptors, pre-defined structural features and any added scaffolds/external descriptors using either automatic or manual selection. In LPDM's automatic descriptor selection, all singletons and non-differentiating descriptors are first removed, and then a *t*- or χ^2 -test is used to evaluate the influence of each descriptor on the continuous or binary response variable, respectively [85]. Then it selects the top 30% of the descriptors according to the χ^2 -test for a binary response variable, or the top and bottom 15% according to the *t*-test for a

⁶ <http://www.leadscope.com/>

continuous response variable. In the manual mode, the model developer selects the preferred descriptors for model building.

After phase 1) descriptor pre-selection, the third and final LPDM model development step starts. In this step, LPDM builds a predictive model using PLS regression for a training set with a continuous response variable and PLR for a binary response variable [85,86]. In the PLS or PLR, the descriptors are used in factors that are extracted and rotated one at a time to maximize the correlation between a principal component and the response variable [85]. The default maximum number of factors is 10 but the model developer can change this maximum or choose a fixed number of factors. In LPDM's automated model building mode, a model is first built using all the phase 1) pre-selected descriptors. During model building, a *k*-fold cross-validation procedure is performed that outputs a predicted residual error sum of square (PRESS) and other statistical measures. The default size of *k* depends on the size of the training set but *k* can also be manually set by the model developer. The descriptors with low loading, low weight, and high residuals in the model are identified, and of these between 5 and 25 are removed. The reduced descriptor set is used in a new model building and cross-validation round. This procedure is repeated until up to 15 predictive models have been built, and among these preliminary models the model with the lowest PRESS is selected as the final model.

It is important to note that LPDM's cross-validation method transfers information such as the selected descriptors from the full model to the smaller cross-validation models. This is therefore in a mathematical sense not a true cross-validation and due to the reuse of information the cross-validation estimates have a tendency to be overoptimistic in its measures on model performance.

Many other software tools for QSAR development exist, both commercially and freely available, including open-source. They use a wide variety of the descriptor sources, descriptor selection methods as well as QSAR modeling algorithms. Examples include SciQSAR, MultiCase CASE Ultra [77] as well as packages in MATLAB⁷, R⁸ and Python's Scikit-learn⁹ [72]. An overview is available from the EU Antares project¹⁰.

2.2.2 The OECD Principles for Validation of QSAR Models

To facilitate the use of QSARs for e.g. regulatory purposes in the context of chemical hazard and risk assessment a need to harmonize the validation of QSAR models arose [1]. At the international workshop 'Regulatory Acceptance of QSARs for Human Health and Environmental Endpoints' held in 2002 in Setubal, Portugal [87], six principles were proposed for assessing the validity of QSAR models [1]. Subsequently, an OECD (Organisation for Economic Co-operation and Development) Expert

⁷ <https://se.mathworks.com/products/matlab.html>

⁸ <https://www.r-project.org/>

⁹ <http://scikit-learn.org/stable/>

¹⁰ <http://www.antares-life.eu/index.php?sec=modellist>

Group assessment of the principles resulted in two of the principles being merged into a single principle. This resulted in the adoption of five OECD principles for QSAR validation in 2004 [1,88,89]. Together, the five OECD principles focus on the scientific validity, i.e. relevance and reliability, of a model [1]. For a QSAR result to be adequate for regulatory use the estimate should be generated by a scientifically valid model that is applicable to the chemical of interest with the necessary level of reliability and whose endpoint is assessed relevant for the regulatory purpose [1]. For regulatory acceptance, the QSAR models and their validation, including the five OECD principles, should be documented in the QSAR Model Reporting Format (QMRF), and the individual QSAR predictions should be documented in the QSAR Prediction Reporting Format (QPRF) [1]. These two documents can be used by the authorities to assess whether the applied model is scientifically valid and fit for purpose, and if the prediction is reliable and adequate enough to be included in a chemical hazard or risk assessment [1]. Guidance on the principles has been described in several documents [1,88–91]. Here is a short introduction and discussion of the OECD QSAR validation principles:

1. A defined endpoint

This principle is intended to ensure clarity and transparency in the endpoint being predicted by the given model. Endpoint refers to any physico-chemical property, biological effect, or environmental parameter that can be measured and modeled. The nature and sources of the experimental data used in the training set have an influence on the reliability of the model. If data originates from multiple sources or varying testing/data analysis protocols, this can affect the model performance as these (small) variations will be built into the model. By providing adequate information on the endpoint, the model user can evaluate if the endpoint and the quality of the underlying data comply with his or her standards for the intended purpose.

2. An unambiguous algorithm

To ensure transparency in the description of the model algorithm with the purpose of having reproducible predictions, the QSAR model should preferably be expressed in the form of an unambiguous algorithm. Full transparency is often not possible when applying a commercial software or very complex model algorithms but in such cases a detailed description of the software and/or modeling process can be given to provide sufficient information for reproducing the model and predictions under the same conditions.

3. A defined applicability domain

A defined AD should be given to describe the limitations of the model in terms of the types of chemical structures, physico-chemical properties and mechanisms of actions for which the model can return reliable predictions. This principle is important to ensure that the QSAR model only makes

interpolations based on the information from its training set. Multiple AD definitions can be applied to the same model depending on its purpose and how reliable predictions the user/developer requires. Generally, a stricter AD results in models with smaller coverage but higher predictive performance as a consequence of excluding less reliable predictions [15,92]. However, this general rule depends on the training set and the method and definition used for AD and in some cases predictions outside the AD can be as accurate as the predictions inside the AD [79,92].

4. Appropriate measures of goodness-of-fit, robustness and predictivity

This principle covers the statistical validation of the QSAR models and the methods are introduced in section 2.2.1.4. In general, two types of statistical information are required to assess the model's goodness-of-fit, robustness and predictive performance: a) an internal performance determined by predicting the training set; and b) an assessment of the model's predictivity of a test set, i.e. a set of chemical structures never seen by the model. The goodness-of-fit serves to provide statistical information for a). The model predictivity statistics for b) can be derived from robust external validation and/or from robust cross-validation that will in addition provide information of model robustness.

5. A mechanistic interpretation

The intent of this principle is to ensure that any identified mechanistic association between descriptors used in the model and the model endpoint are documented. A mechanistic interpretation can further strengthen the confidence in the model established based on the previous four principles. It is not always possible to provide a mechanistic interpretation of a QSAR model however, and it is furthermore important to keep in mind that even if a strong correlation is found between descriptor(s) and the response variable this does not imply that there is causality.

2.2.3 The Danish (Q)SAR Database

The current version of the Danish (Q)SAR Database (<http://qsar.food.dtu.dk/>) was released in November 2015 and replaced the previous version from 2004. It is a free, online database with structural information, QSAR predictions, and in some cases experimental results, for ~640,000 discrete organic chemical substances [78]. It is developed and maintained at the Technical University of Denmark (DTU) with support from the Danish Environmental Protection Agency (EPA) and Nordic Council of Ministers. More than 200 global QSAR models have been applied for around 45 endpoints covering physico-chemical properties, molecular mechanisms including mutagenesis and receptor binding, to *in vivo* and clinical endpoints. Most endpoints have been modeled in three different commercial QSAR systems: LPDM, Scimatics SciQSAR and MultiCASE® CASE Ultra [77]. The individual predictions from each system as well as a battery prediction call integrating the three predictions are

available. QMRFs for all the applied models are provided. The online database is capable of doing complex search queries, including substructure, similarity and property searches or combinations of these. The predictions in the Danish (Q)SAR Database can be used in for example screening, profiling and prioritization by industry, academia, agencies and NGOs. The database is dynamic and predictions from new models will continuously be added, for example predictions from the LPDM models developed in this PhD project. All predictions in the Danish (Q)SAR Database will be incorporated into the OECD (Q)SAR Toolbox [93], where the predictions together with other information can be used in constructing chemical categories for grouping and read-across purposes.

Currently under development is a ‘sister-site’ to the Danish (Q)SAR Database. Here the in-house LPDM models from the Danish (Q)SAR Database, including the models developed in this PhD project, will be made available for free prediction of user-submitted structures. Besides predictions of structures not in the Danish (Q)SAR Database, users will have access to more prediction details such as analog structures from training sets and model structural features used to produce the predictions.

2.2.4 Application of QSAR

QSARs are used in multiple chemical research areas such as drug discovery and toxicology [2], and they are among other things applied to:

- increase the amount of (toxicological) information on chemicals
- help prioritize and rank chemicals/drugs for further testing or evaluation [94]
- help the (medical) chemist optimize structures to a given target [31]
- help design safer substitution chemicals
- contribute to the reduction and replacement of animal testing [95]

Furthermore, since a QSAR model averages over all the closest analogs in the training set, it is possible for an individual model estimate to be more accurate than an individual experimental measurement, and QSARs can in some cases cause identification of chemicals with erroneous experimental results [1,12,22]. Below are some examples on the application of QSAR.

2.2.4.1 QSAR in Regulations

The regulatory interest and use of QSAR is steadily increasing as they hold the potential to help fill the large gaps in toxicological information of the many thousands of man-made chemicals queued for risk assessment and classification and labeling [2,79,95–101]. Furthermore, QSAR results provide additional mechanistic information useful in for example grouping of chemicals into categories for read-across and improve evaluation of existing test data [1]. Multiple examples on the use of QSAR for replacement or supplement of experimental data in regulatory contexts exist for physico-

chemical properties, environmental fate parameters and ecotoxicological endpoints [1,94,102–105]. For human health effects, however, the application of QSARs is still in its early phase [103] and has primarily been used as a supplement to experimental data and for groupings and prioritization purposes [1,94]. Facing forward, QSARs are expected to be used increasingly for direct replacement of test data as the experience in and acceptance of QSARs and their predictions become more widespread within the regulatory community [1,95].

Examples of regulatory implementation of QSARs can be found in EU's chemicals regulation, REACH [101], and the International Council for Harmonisation (ICH) M7 guideline [100]. Briefly, the ICH M7 guideline describes the approach to identify, categorize and control DNA reactive, mutagenic impurities in pharmaceutical products to limit the potential carcinogenic risk from such impurities [100,106]. Here (Q)SAR predictions from two complementary QSAR methodologies, i.e. a statistical-based and an expert rule-based, followed by expert review may be used for classification of drug impurities in case of missing experimental data. The absence of structural alerts from the two complementary (Q)SAR methodologies is sufficient to conclude that the impurity is of no mutagenic concern, and no further testing is recommended [100].

2.2.4.2 QSAR in Screening and Prioritization

QSAR models are useful tools for screening and prioritization of chemicals for further testing. For example QSARs can be used in a tiered screening approach where the most problematic chemicals or the most promising drug candidates based on QSAR predictions are prioritized for further *in vitro* and/or *in vivo* testing [1,62,94,107].

The Danish EPA has for around two decades supported a number of activities on research and development as well application of QSARs for screening in regulatory contexts. For example, the Danish EPA together with QSAR researchers from the National Food Institute, DTU, has since 2001 published four versions of the Advisory list for self-classification of dangerous substances [108–111]. In these projects, QSAR predictions for a number of endpoints of relevance for acute oral toxicity, skin sensitization and irritation, mutagenicity, carcinogenicity, reproductive toxicity (i.e. possible harm to the unborn child) and danger to the aquatic environment were used to make advisory classifications for ~33,835 EINECS (European Inventory of Existing Commercial Chemical Substances) substances according to the CLP-regulation (classification, labelling and packaging of substances and mixtures) criteria [96,109]. A second example is a Danish EPA supported project from 2013 that describes the use of QSAR to identify potential CMR (carcinogenic, mutagenic or toxic to reproduction) REACH substances according to the CLP-regulation [112]. Screening results from

QSARs have also recently been used by the Danish EPA for grouping a number of brominated flame retardants [113].

2.2.4.3 QSAR in Early Drug Development

Because of the time and cost demanding process of bringing a new drug to the market and the high attrition rate [114,115], the pharmaceutical industry is striving towards implementation of technologies that can optimize the process [116]. The application of *in silico* methods for ligand-based virtual screening (LBVS), including QSAR models, has become a routine tool in drug design and early drug discovery phases in some pharmaceutical companies [117,118]. QSARs are used for fast screening of large sets of virtual small-molecule drug candidates to identify activity towards the drug target as well as toxicological properties [62,119]. QSARs are also used by the medical chemist to identify chemical features involved in the drug target activity and this information can be used for optimizing and isolating drug candidates [31,118,120].

2.2.4.4 QSAR in Hypothesis Generation

If information for two or more different biological endpoints is available for a big and diverse set of chemicals, statistical correlations between the results from the endpoints can be calculated, and if a significant correlation is found this may be an indication of a biological association between the endpoints. The correlations can be performed using different methods such as univariate or multivariate data analysis. A number of papers using univariate data analysis for correlation studies between results from an array of HTS *in vitro* and an *in vivo* endpoint have been published [121] and can help researchers generate new hypotheses on associations between molecular mechanism(s) and effects at the organ/organism level. This data-driven inductive and holistic approach for hypothesis generation [122] holds the limitation of restrictions in the number of overlapping structures having experimental results in the studied endpoints. With QSAR models it is possible to generate information for multiple biological endpoints for a large and structurally diverse set of structures, which can then be used for performing statistical correlations [36,123] and generating new hypotheses. It is important to keep in mind that the associations are purely statistical and the generated biological hypotheses will need to be tested by applying other techniques.

References

- [1] ECHA, Guidance on information requirements and chemical safety assessment - Chapter R.6: QSARs and grouping of chemicals, (2008).
https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf (accessed March 16, 2017).
- [2] A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I.I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y.C. Martin, R. Todeschini, V. Consonni, V.E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, QSAR Modeling: Where Have You Been? Where Are You Going To?, *J. Med. Chem.* 57 (2014) 4977–5010. doi:10.1021/jm4004285.
- [3] J.C. Dearden, M.D. Barratt, R. Benigni, W. Douglas, R.D. Combes, M.T.D. Cronin, P.N. Judson, M.P. Payne, A.M. Richard, M. Tichy, A.P. Worth, J.J. Yourick, The Development and Validation of Expert Systems for Predicting Toxicity, *Altern. to Lab. Anim.* 25 (1997) 223–252.
- [4] F. Baum, Zur Theorie der Alkoholnarkose, *Arch. Für Exp. Pathol. Und Pharmakologie.* 42 (1899) 119–137. doi:10.1007/BF01834480.
- [5] R.L. Lipnick, Hans Horst Meyer and the lipoid theory of narcosis, *Trends Pharmacol. Sci.* 10 (1989) 265–269. doi:10.1016/0165-6147(89)90025-4.
- [6] H. Meyer, Zur Theorie der Alkoholnarkose, *Arch. Für Exp. Pathol. Und Pharmakologie.* 42 (1899) 109–118. doi:10.1007/BF01834479.
- [7] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, A practical overview of quantitative structure-activity relationship, *EXCLI J.* 8 (2009) 74–88.
- [8] T. Fujita, J. Iwasa, C. Hansch, A New Substituent Constant, π , Derived from Partition Coefficients, *J. Am. Chem. Soc.* 86 (1964) 5175–5180. doi:10.1021/ja01077a028.
- [9] C. Hansch, E. Deutsch, The structure-activity relationship in amides inhibiting photosynthesis, *Biochim. Biophys. Acta.* 5 (1966) 381–391.
- [10] C. Hansch, P.P. Maloney, T. Fujita, R.M. Muir, Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients, *Nature.* 194 (1962) 178–180. doi:10.1038/194178b0.
- [11] C. Hansch, R.M. Muir, T. Fujita, P.P. Maloney, F. Geiger, M. Streich, The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients, *J. Am. Chem. Soc.* 85 (1963) 2817–2824. doi:10.1021/ja00901a033.
- [12] A. Tropsha, Best Practices for QSAR Model Development, Validation, and Exploitation, *Mol. Inform.* 29 (2010) 476–488. doi:10.1002/minf.201000061.
- [13] OECD, Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment, (2005).
[http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2005\)14&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2005)14&doclanguage=en) (accessed March 20, 2017).
- [14] A.P. Worth, A. Bassan, J. De Bruijn, A. Gallegos Saliner, T. Netzeva, M. Pavan, G. Patlewicz, I. Tsakovska, S. Eisenreich, The role of the European Chemicals Bureau in promoting the regulatory use of (Q)SAR methods†, *SAR QSAR Environ. Res.* 18 (2007) 111–125. doi:10.1080/10629360601054255.
- [15] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R.

- Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. Van De Sandt, W. Tong, G. Veith, C. Yang, Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships, *Altern. to Lab. Anim.* 33 (2005) 155–173.
- [16] D. Fourches, E. Muratov, A. Tropsha, Curation of chemogenomics data, *Nat. Chem. Biol.* 11 (2015) 535–535. doi:10.1038/nchembio.1881.
- [17] M. Gütlein, C. Helma, A. Karwath, S. Kramer, A Large-Scale Empirical Evaluation of Cross-Validation and External Test Set Validation in (Q)SAR, *Mol. Inform.* 32 (2013) 516–528. doi:10.1002/minf.201200134.
- [18] R. Huang, M. Xia, S. Sakamuru, J. Zhao, S.A. Shahane, M. Attene-Ramos, T. Zhao, C.P. Austin, A. Simeonov, Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization, *Nat. Commun.* 7 (2016) 10425. doi:10.1038/ncomms10425.
- [19] B.L. Ingle, B.C. Veber, J.W. Nichols, R. Tornero-Velez, Informing the Human Plasma Protein Binding of Environmental Chemicals by Machine Learning in the Pharmaceutical Space: Applicability Domain and Limits of Predictability, *J. Chem. Inf. Model.* 56 (2016) 2243–2252. doi:10.1021/acs.jcim.6b00291.
- [20] F.P. Steinmetz, S.J. Enoch, J.C. Madden, M.D. Nelms, N. Rodriguez-Sanchez, P.H. Rowe, Y. Wen, M.T.D. Cronin, Methods for assigning confidence to toxicity data with multiple values — Identifying experimental outliers, *Sci. Total Environ.* 482–483 (2014) 358–365. doi:10.1016/j.scitotenv.2014.02.115.
- [21] D. Fourches, E. Muratov, A. Tropsha, Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation, *J. Chem. Inf. Model.* 56 (2016) 1243–1252. doi:10.1021/acs.jcim.6b00129.
- [22] D. Fourches, E. Muratov, A. Tropsha, Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research, *J. Chem. Inf. Model.* 50 (2010) 1189–1204. doi:10.1021/ci100176x.
- [23] K. Mansouri, C.M. Grulke, A.M. Richard, R.S. Judson, A.J. Williams, An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling, *SAR QSAR Environ. Res.* 27 (2016) 911–937. doi:10.1080/1062936X.2016.1253611.
- [24] E. Anderson, G.D. Veith, D. Weininger, SMILES: A line notation and computerized interpreter for chemical structures, 1987. <https://nepis.epa.gov/Exe/ZyNET.exe/2000CAUR.TXT?ZyActionD=ZyDocument&Client=EPA&Index=1986+Thru+1990&Docs=&Query=&Time=&EndTime=&SearchMethod=1&TocRestrict=n&Toc=&TocEntry=&QField=&QFieldYear=&QFieldMonth=&QFieldDay=&IntQFieldOp=0&ExtQFieldOp=0&XmlQuery=> (accessed February 20, 2017).
- [25] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Model.* 28 (1988) 31–36. doi:10.1021/ci00057a005.
- [26] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, *J. Chem. Inf. Model.* 29 (1989) 97–101. doi:10.1021/ci00062a008.
- [27] A. Dalby, J.G. Nourse, W.D. Hounshell, A.K.I. Gushurst, D.L. Grier, B.A. Leland, J. Laufer, Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited, *J. Chem. Inf. Model.* 32 (1992) 244–255. doi:10.1021/ci00007a012.
- [28] S. Dastmalchi, M. Hamzeh-Mivehroud, K. Asadpour-Zeynali, Comparison of Different 2D and 3D-QSAR Methods on Activity Prediction of Histamine H3 Receptor Antagonists, *Iran. J. Pharm. Res. IJPR.* 11 (2012) 97–108. <http://www.ncbi.nlm.nih.gov/pubmed/25317190> (accessed January 10, 2017).

- [29] Y. Fang, Y. Lu, X. Zang, T. Wu, X. Qi, S. Pan, X. Xu, 3D-QSAR and docking studies of flavonoids as potent *Escherichia coli* inhibitors, *Sci. Rep.* 6 (2016) 23634. doi:10.1038/srep23634.
- [30] O. Mekenyan, N. Nikolova, P. Schmieder, Dynamic 3D QSAR techniques: applications in toxicology, *J. Mol. Struct.* 622 (2003) 147–165. doi:10.1016/S0166-1280(02)00625-5.
- [31] J. Verma, V. Khedkar, E. Coutinho, 3D-QSAR in Drug Design - A Review, *Curr. Top. Med. Chem.* 10 (2010) 95–115. doi:10.2174/156802610790232260.
- [32] J. Polanski, Receptor Dependent Multidimensional QSAR for Modeling Drug - Receptor Interactions, *Curr. Med. Chem.* 16 (2009) 3243–3257. doi:10.2174/092986709788803286.
- [33] D. Young, T. Martin, R. Venkatapathy, P. Harten, Are the Chemical Structures in Your QSAR Correct?, *QSAR Comb. Sci.* 27 (2008) 1337–1345. doi:10.1002/qsar.200810084.
- [34] A.M. Richard, R.S. Judson, K.A. Houck, C.M. Grulke, P. Volarath, I. Thillainadarajah, C. Yang, J. Rathman, M.T. Martin, J.F. Wambaugh, T.B. Knudsen, J. Kancherla, K. Mansouri, G. Patlewicz, A.J. Williams, S.B. Little, K.M. Crofton, R.S. Thomas, ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology, *Chem. Res. Toxicol.* 29 (2016) 1225–1251. doi:10.1021/acs.chemrestox.6b00135.
- [35] J.D. Walker, J. Jaworska, M.H.I. Comber, T.W. Schultz, J.C. Dearden, Guidelines for developing and using Quantitative Structure-Activity Relationships, *Environ. Toxicol. Chem.* 22 (2003) 1653–1665. doi:10.1897/01-627.
- [36] S.A. Rosenberg, M. Xia, R. Huang, N.G. Nikolov, E.B. Wedebye, M. Dybdahl, QSAR development and profiling of 72,524 REACH substances for PXR activation and CYP3A4 induction, *Comput. Toxicol.* 1 (2017) 39–48. doi:10.1016/j.comtox.2017.01.001.
- [37] G.M. Maggiora, On Outliers and Activity Cliffs - Why QSAR Often Disappoints, *J. Chem. Inf. Model.* 46 (2006) 1535–1535. doi:10.1021/ci060117s.
- [38] M.T.D. Cronin, T.W. Schultz, Pitfalls in QSAR, *J. Mol. Struct.* 622 (2003) 39–51. doi:10.1016/S0166-1280(02)00616-4.
- [39] P.P. Roy, J.T. Leonard, K. Roy, Exploring the impact of size of training sets for the development of predictive QSAR models, *Chemom. Intell. Lab. Syst.* 90 (2008) 31–42. doi:10.1016/j.chemolab.2007.07.004.
- [40] T.M. Martin, P. Harten, D.M. Young, E.N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling?, *J. Chem. Inf. Model.* 52 (2012) 2570–2578. doi:10.1021/ci300338w.
- [41] A. Nandy, S. Kar, K. Roy, Development of classification- and regression-based QSAR models and in silico screening of skin sensitisation potential of diverse organic chemicals, *Mol. Simul.* 40 (2014) 261–274. doi:10.1080/08927022.2013.801076.
- [42] A. Golbraikh, A. Tropsha, Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection, *Mol. Divers.* 5 (2000) 231–243. doi:10.1023/A:1021372108686.
- [43] S.J. Capuzzi, R. Politi, O. Isayev, S. Farag, A. Tropsha, QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays, *Front. Environ. Sci.* 4 (2016) 1–7. doi:10.3389/fenvs.2016.00003.
- [44] A. V. Zakharov, M.L. Peach, M. Sitzmann, M.C. Nicklaus, QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem, *J. Chem. Inf. Model.* 54 (2014) 705–712. doi:10.1021/ci400737s.
- [45] J.J. Chen, C. A. Tsai, J.F. Young, R.L. Kodell, Classification ensembles for unbalanced class sizes

- in predictive toxicology, *SAR QSAR Environ. Res.* 16 (2005) 517–529. doi:10.1080/10659360500468468.
- [46] L. Breiman, Bagging Predictors, *Mach. Learn.* 24 (1996) 123–140. doi:10.1023/A:1018054314350.
- [47] P. Lee, Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets, *Int. J. Environ. Res. Public Health*. 11 (2014) 9776–9789. doi:10.3390/ijerph110909776.
- [48] N. Japkowicz, Learning from Imbalanced Data Sets: A Comparison of Various Strategies *, (2000). <https://pdfs.semanticscholar.org/1af9/6acae07b1e141f98f3df973eaf9e0a9226fb.pdf> (accessed March 14, 2017).
- [49] Q. Zang, D.M. Rotroff, R.S. Judson, Binary Classification of a Large Collection of Environmental Chemicals from Estrogen Receptor Assays by Quantitative Structure–Activity Relationship and Machine Learning Methods, *J. Chem. Inf. Model.* 53 (2013) 3244–3261. doi:10.1021/ci400527b.
- [50] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of MDL Keys for Use in Drug Discovery, *J. Chem. Inf. Comput. Sci.* 42 (2002) 1273–1280. doi:10.1021/ci010132r.
- [51] G. Roberts, G.J. Myatt, W.P. Johnson, K.P. Cross, P.E. Blower, LeadScope † : Software for Exploring Large Sets of Screening Data, *J. Chem. Inf. Comput. Sci.* 40 (2000) 1302–1314. doi:10.1021/ci0000631.
- [52] K.P. Cross, G. Myatt, C. Yang, M.A. Fligner, J.S. Verducci, P.E. Blower, Finding Discriminating Structural Features by Reassembling Common Building Blocks, *J. Med. Chem.* 46 (2003) 4770–4775. doi:10.1021/jm0302703.
- [53] Daylight, 6. Fingerprints - Screening and Similarity, (2017). <http://www.daylight.com/dayhtml/doc/theory/theory.finger.html> (accessed March 14, 2017).
- [54] M. Gütlein, S. Kramer, Filtered circular fingerprints improve either prediction or runtime performance while retaining interpretability, *J. Cheminform.* 8 (2016) 60. doi:10.1186/s13321-016-0173-z.
- [55] P. Jaccard, Etude de la distribution florale dans une portion des Alpes et du Jura, *Bull. La Soc. Vaudoise Des Sci. Nat.* 37 (1901) 547–579. doi:10.5169/seals-266450.
- [56] Danishuddin, A.U. Khan, Descriptors and their selection methods in QSAR analysis: paradigm for drug design, *Drug Discov. Today*. 21 (2016) 1291–1302. doi:10.1016/j.drudis.2016.06.013.
- [57] J. Dong, D.-S. Cao, H.-Y. Miao, S. Liu, B.-C. Deng, Y.-H. Yun, N.-N. Wang, A.-P. Lu, W.-B. Zeng, A.F. Chen, ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation, *J. Cheminform.* 7 (2015) 60. doi:10.1186/s13321-015-0109-z.
- [58] P. Labute, A widely applicable set of descriptors, *J. Mol. Graph. Model.* 18 (2000) 464–477. doi:10.1016/S1093-3263(00)00068-1.
- [59] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, *J. Comput. Chem.* 32 (2011) 1466–1474. doi:10.1002/jcc.21707.
- [60] H. Hotelling, Analysis of a complex of statistical variables into principal components, *Warwick York Inc.* (1933). <http://hdl.handle.net/2027/wu.89097139406> (accessed February 17, 2017).
- [61] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, *Philos. Mag. Ser. 6*. 2 (1901) 559–572. doi:10.1080/14786440109462720.
- [62] M. Danishuddin, A.U. Khan, Structure based virtual screening to discover putative drug

- candidates: Necessary considerations and successful case studies, *Methods*. 71 (2015) 135–145. doi:10.1016/j.ymeth.2014.10.019.
- [63] M. Goodarzi, B. Dejaergher, Y. Vander Heiden, Feature Selection Methods in QSAR Studies, *J. AOAC Int.* 95 (2012) 636–650.
<http://www.ingentaconnect.com/content/aoac/jaoac/2012/00000095/00000003/art00009>.
- [64] M. Shahlaei, Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study, *Chem. Rev.* 113 (2013) 8093–8103. doi:10.1021/cr3004339.
- [65] P. Smialowski, D. Frishman, S. Kramer, Pitfalls of supervised feature selection, *Bioinformatics*. 26 (2010) 440–443. doi:10.1093/bioinformatics/btp621.
- [66] S.P. Niculescu, Artificial neural networks and genetic algorithms in QSAR, *J. Mol. Struct.* 622 (2003) 71–83. doi:10.1016/S0166-1280(02)00619-X.
- [67] R. Judson, F. Elloumi, R.W. Setzer, Z. Li, I. Shah, A comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model, *BMC Bioinformatics*. 9 (2008). doi:10.1186/1471-2105-9-241.
- [68] J.G. Topliss, Utilization of operational schemes for analog synthesis in drug design, *J. Med. Chem.* 15 (1972) 1006–1011. doi:10.1021/jm00280a002.
- [69] P. Liu, W. Long, Current Mathematical Methods Used in QSAR/QSPR Studies, *Int. J. Mol. Sci.* 10 (2009) 1978–1998. doi:10.3390/ijms10051978.
- [70] J.V. Kringelum, Pharmacology profiling of chemicals and proteins, (2014).
[http://orbit.dtu.dk/en/publications/pharmacology-profiling-of-chemicals-and-proteins\(68307564-5fd4-48a3-b38c-8e60a43b058a\).html](http://orbit.dtu.dk/en/publications/pharmacology-profiling-of-chemicals-and-proteins(68307564-5fd4-48a3-b38c-8e60a43b058a).html) (accessed March 14, 2017).
- [71] G.W. Kauffman, P.C. Jurs, QSAR and k-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors, *J. Chem. Inf. Comput. Sci.* 41 (2001) 1553–1560. doi:10.1021/ci010073h.
- [72] A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications, *Drug Discov. Today*. 20 (2015) 318–331. doi:10.1016/j.drudis.2014.10.012.
- [73] B. Chen, R.P. Sheridan, V. Hornak, J.H. Voigt, Comparison of Random Forest and Pipeline Pilot Naïve Bayes in Prospective QSAR Predictions, *J. Chem. Inf. Model.* 52 (2012) 792–803. doi:10.1021/ci200615h.
- [74] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, *J. Chem. Inf. Comput. Sci.* 43 (2003) 1947–1958. doi:10.1021/ci034160g.
- [75] A. Roncaglioni, N. Piclin, M. Pintore, E. Benfenati, Binary classification models for endocrine disrupter effects mediated through the estrogen receptor[†], *SAR QSAR Environ. Res.* 19 (2008) 697–733. doi:10.1080/10629360802550606.
- [76] E. Pourbasheer, R. Aalizadeh, M.R. Ganjali, QSAR study of CK2 inhibitors by GA-MLR and GA-SVM methods, *Arab. J. Chem.* (2015). doi:10.1016/j.arabjc.2014.12.021.
- [77] QSAR, User Manual for the Danish (Q)SAR Database, (2015).
http://qsar.db.food.dtu.dk/Danish_QSAR_Database_Draft_User_manual.pdf (accessed March 28, 2017).
- [78] QSARDB, Danish (Q)SAR Database, (2015). <http://qsar.food.dtu.dk/> (accessed March 14, 2017).
- [79] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A.M. Richard, C.M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E.

- Muratov, E.B. Wedebye, F. Grisoni, G.F. Mangiatordi, G.M. Incisivo, H. Hong, H.W. Ng, I. V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N.G. Nikolov, O. Nicolotti, P.L. Andersson, Q. Zang, R. Politi, R.D. Beger, R. Todeschini, R. Huang, S. Farag, S.A. Rosenberg, S. Slavov, X. Hu, R.S. Judson, CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ. Health Perspect.* 124 (2016) 1023–1033. doi:10.1289/ehp.1510267.
- [80] J.A. Cooper II, R. Saracci, P. Cole, Describing the validity of carcinogen screening tests, *Br. J. Cancer.* 39 (1979) 87–89.
- [81] P. Gramatica, Principles of QSAR models validation: internal and external, *QSAR Comb. Sci.* 26 (2007) 694–701. doi:10.1002/qsar.200610151.
- [82] P. Gramatica, External Evaluation of QSAR Models, in Addition to Cross-Validation: Verification of Predictive Capability on Totally New Chemicals, *Mol. Inform.* 33 (2014) 311–314. doi:10.1002/minf.201400030.
- [83] D.M. Hawkins, The Problem of Overfitting, *J. Chem. Inf. Comput. Sci.* 44 (2004) 1–12. doi:10.1021/ci0342472.
- [84] N. Nikolov, V. Grancharov, G. Stoyanova, T. Pavlov, O. Mekenyan, Representation of Chemical Information in OASIS Centralized 3D Database for Existing Chemicals, *J. Chem. Inf. Model.* 46 (2006) 2537–2551. doi:10.1021/ci060142y.
- [85] L.G. Valerio, C. Yang, K.B. Arvidson, N.L. Kruhlak, A structural feature-based computational approach for toxicology predictions, *Expert Opin. Drug Metab. Toxicol.* 6 (2010) 505–518. doi:10.1517/17425250903499286.
- [86] C. Yang, K. Cross, G.J. Myatt, P.E. Blower, J.F. Rathman, Building Predictive Models for Protein Tyrosine Phosphatase 1B Inhibitors Based on Discriminating Structural Features by Reassembling Medicinal Chemistry Building Blocks, *J. Med. Chem.* 47 (2004) 5984–5994. doi:10.1021/jm0497242.
- [87] J.S. Jaworska, M. Comber, C. Auer, C.J. Van Leeuwen, Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints, *Environ. Health Perspect.* 111 (2003) 1358–1360. doi:10.1289/ehp.5757.
- [88] OECD, OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationships models, (2004) 1–2. www.oecd.org/dataoecd/33/37/37849783.pdf (accessed March 13, 2017).
- [89] OECD, The report from the expert group on (quantitative) structure-activity relationships [(Q)SARs] on the principles for the validation of (Q)SARs, (2004). [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2004\)24&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2004)24&doclanguage=en) (accessed March 13, 2017).
- [90] OECD, Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models, (2007). [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono\(2007\)2](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2) (accessed March 14, 2017).
- [91] A.P. Worth, A. Bassan, A. Gallegos, T.. Netzeva, G. Patlewicz, M. Pavan, I. Tsakovska, M. Vracko, The characterisation of (Quantitative) Structure-Activity Relationships: Preliminary guidance, *ECB Rep. EUR 21866 Eur. Commission, Jt. Res. Cent.* (2005).
- [92] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of Different Approaches to Define the Applicability Domain of QSAR Models, *Molecules.* 17 (2012) 4791–4810. doi:10.3390/molecules17054791.

- [93] OECD, The OECD QSAR Toolbox, (2015). <http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm> (accessed March 14, 2017).
- [94] C.L. Russom, R.L. Breton, J.D. Walker, S.P. Bradbury, An overview of the use of Quantitative Structure-Activity Relationships for ranking and prioritizing large chemical inventories for environmental risk assessments, *Environ. Toxicol. Chem.* 22 (2003) 1810–1821. doi:10.1897/01-194.
- [95] K. Stanton, F.H. Kruszewski, Quantifying the benefits of using read-across and in silico techniques to fulfill hazard data requirements for chemical categories, *Regul. Toxicol. Pharmacol.* 81 (2016) 250–259. doi:10.1016/j.yrtph.2016.09.004.
- [96] CLP, Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, (2008). <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008R1272&from=EN> (accessed March 16, 2017).
- [97] EC SCCS, The SCCS's notes of guidance for the testing of cosmetic ingredients and their safety evaluation, (2016). http://ec.europa.eu/health/scientific_committees/consumer_safety/docs/sccs_o_190.pdf (accessed March 16, 2017).
- [98] EFSA, Guidance on the establishment of the residue definition for dietary risk assessment, *EFSA J.* 14 (2016). doi:10.2903/j.efsa.2016.4549.
- [99] EU, Regulation (EU) No 528/2012 of the European Parliament and of the Council 22 May 2012 concerning the making available on the market and use of biocidal products, (2012). <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012R0528&from=EN> (accessed March 16, 2017).
- [100] ICH, M7 Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk, ICH Harmon. Tripart. Guidel. (2015) 35. <https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM347725.pdf>.
- [101] REACH, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), (2006). <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02006R1907-20161011&from=EN>.
- [102] EC TGD, Technical Guidance Document on Risk Assessment, (2003). https://echa.europa.eu/documents/10162/16960216/tgdpart2_2ed_en.pdf (accessed March 1, 2017).
- [103] S. Gutsell, P. Russell, The role of chemistry in developing understanding of adverse outcome pathways and their application in risk assessment, *Toxicol. Res. (Camb)*. 2 (2013) 299–307. doi:10.1039/c3tx50024a.
- [104] US EPA, TSCA New Chemicals Program (NCP) Chemical Categories, (2010). https://www.epa.gov/sites/production/files/2014-10/documents/ncp_chemical_categories_august_2010_version_0.pdf (accessed March 16, 2017).
- [105] M.T.D. Cronin, J.S. Jaworska, J.D. Walker, M.H.I. Comber, C.D. Watts, A.P. Worth, Use of QSARs in International Decision-Making Frameworks to Predict Health Effects of Chemical Substances, *Environ. Health Perspect.* 111 (2002) 1391–1401. doi:10.1289/ehp.5760.
- [106] A. Amberg, L. Beilke, J. Bercu, D. Bower, A. Brigo, K.P. Cross, L. Custer, K. Dobo, E. Dowdy, K.A. Ford, S. Glowienke, J. Van Gompel, J. Harvey, C. Hasselgren, M. Honma, R. Jolly, R. Kemper,

- M. Kenyon, N. Kruhlak, P. Leavitt, S. Miller, W. Muster, J. Nicolette, A. Plaper, M. Powley, D.P. Quigley, M.V. Reddy, H.-P. Spirk, L. Stavitskaya, A. Teasdale, S. Weiner, D.S. Welch, A. White, J. Wichard, G.J. Myatt, Principles and procedures for implementation of ICH M7 recommended (Q)SAR analyses, *Regul. Toxicol. Pharmacol.* 77 (2016) 13–24. doi:10.1016/j.yrtph.2016.02.004.
- [107] EDSP21 Work Plan, The Incorporation of In Silico Models and In Vitro High Throughput Assays in the Endocrine Disruptor Screening Program (EDSP) for Prioritization and Screening, (2011). https://www.epa.gov/sites/production/files/2015-07/documents/edsp21_work_plan_summary_overview_final.pdf (accessed March 13, 2017).
- [108] Danish EPA, Report on the Advisory list for selfclassification of dangerous substances - Environmental Project No. 636, 2001. <http://www2.mst.dk/Udgiv/publications/2001/87-7944-694-9/pdf/87-7944-695-7.pdf> (accessed March 21, 2017).
- [109] J.R. Niemelä, E.B. Wedebye, N.G. Nikolov, G.E. Jensen, T. Ringsted, F. Ingerslev, H. Tyle, C. Ihlemann, The Advisory list for self- classification of dangerous substances - Environmental Project No. 1351, 2010. <http://www2.mst.dk/udgiv/publications/2010/978-87-92708-58-8/pdf/978-87-92708-59-5.pdf> (accessed March 16, 2017).
- [110] J.R. Niemelä, E.B. Wedebye, N.G. Nikolov, G.E. Jensen, T. Ringsted, F. Ingerslev, H. Tyle, C. Ihlemann, The Advisory list for self- classification of dangerous substances - Environmental Project No. 1322, 2010. <http://www2.mst.dk/udgiv/publications/2010/978-87-92617-64-4/pdf/978-87-92617-65-1.pdf> (accessed March 21, 2017).
- [111] J.R. Niemelä, E.B. Wedebye, N.G. Nikolov, G.E. Jensen, T. Ringsted, F. Ingerslev, H. Tyle, C. Ihlemann, The Advisory list for self- classification of dangerous substances - Environmental Project No. 1303, 2009. <http://www2.mst.dk/udgiv/publications/2009/978-87-92548-56-6/pdf/978-87-92548-57-3.pdf> (accessed March 21, 2017).
- [112] E.B. Wedebye, J.R. Niemelä, N.G. Nikolov, M. Dybdahl, Use of QSAR to identify potential CMR substances of relevance under the REACH regulation, 2013. <http://www2.mst.dk/Udgiv/publications/2013/09/978-87-93026-48-3.pdf> (accessed March 1, 2017).
- [113] Danish EPA, Category approach for selected brominated flame retardants, 2016. <http://www2.mst.dk/Udgiv/publications/2016/07/978-87-93435-90-2.pdf> (accessed February 17, 2017).
- [114] I. Kola, J. Landis, Opinion: Can the pharmaceutical industry reduce attrition rates?, *Nat. Rev. Drug Discov.* 3 (2004) 711–716. doi:10.1038/nrd1470.
- [115] H. Olson, G. Betton, D. Robinson, K. Thomas, A. Monro, G. Kolaja, P. Lilly, J. Sanders, G. Sipes, W. Bracken, M. Dorato, K. Van Deun, P. Smith, B. Berger, A. Heller, Concordance of the Toxicity of Pharmaceuticals in Humans and in Animals, *Regul. Toxicol. Pharmacol.* 32 (2000) 56–67. doi:10.1006/rtph.2000.1399.
- [116] R.D. Clark, W. Liang, A.C. Lee, M.S. Lawless, R. Fraczekiewicz, M. Waldman, Using beta binomials to estimate classification uncertainty for ensemble models, *J. Cheminform.* 6 (2014) 1–19. doi:10.1186/1758-2946-6-34.
- [117] C.-H. Lee, H.-C. Huang, H.-F. Juan, Reviewing Ligand-Based Rational Drug Design: The Search for an ATP Synthase Inhibitor, *Int. J. Mol. Sci.* 12 (2011) 5304–5318. doi:10.3390/ijms12085304.
- [118] N. Ogihara, Drawing Out Drugs, *Mod. Drug Discov.* 6 (2003) 28–31.
- [119] A. Roncaglioni, A.A. Toropov, A.P. Toropova, E. Benfenati, In silico methods to predict drug toxicity, *Curr. Opin. Pharmacol.* 13 (2013) 802–806. doi:10.1016/j.coph.2013.06.001.

- [120] R.D. Cramer, The inevitable QSAR renaissance, *J. Comput. Aided. Mol. Des.* 26 (2012) 35–38. doi:10.1007/s10822-011-9495-0.
- [121] R. Kavlock, K. Chandler, K. Houck, S. Hunter, R. Judson, N. Kleinstreuer, T. Knudsen, M. Martin, S. Padilla, D. Reif, A. Richard, D. Rotroff, N. Sipes, D. Dix, Update on EPA's ToxCast Program: Providing High Throughput Decision Support Tools for Chemical Risk Management, *Chem. Res. Toxicol.* 25 (2012) 1287–1302. doi:10.1021/tx3000939.
- [122] D.B. Kell, S.G. Oliver, Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era, *BioEssays*. 26 (2004) 99–105. doi:10.1002/bies.10385.
- [123] M. Dybdahl, N.G. Nikolov, E.B. Wedebye, S.Ó. Jónsdóttir, J.R. Niemelä, QSAR model for human pregnane X receptor (PXR) binding: Screening of environmental chemicals and correlations with genotoxicity, endocrine disruption and teratogenicity, *Toxicol. Appl. Pharmacol.* 262 (2012) 301–309. doi:10.1016/j.taap.2012.05.008.

2.3 Regulatory Toxicology

Toxicology, from the ancient Greek words *toxikos* (“poisonous”) and *logia* (“study of”), is the study of adverse effects of chemical substances on living organisms and was founded as a research field by Paracelsus (1493-1541 CE) [1]. Today, it applies theories and methods from multiple disciplines such as biology, biochemistry and computer science to identify a chemical’s potential adverse effects, which is influenced by factors such as dosage, time and route of exposure, properties of the exposed organism (sex, age, health, etc.) as well as other environmental factors (simultaneous exposure to other chemicals, temperature, etc.).

The production and diversity of man-made chemicals applied in industry, agriculture, war and consumer products are steadily increasing, and imprint of such chemicals can be found all over the world today [2,3]. Because of their potential adverse impact on human health and the environment, there is increasing concern about the safety of the chemicals in our surroundings. Chemicals are subject to different national and international chemical regulations that require different levels of toxicity information depending on their production volume, use, etc. [4–8]. A chemical risk assessment combines information from hazard identification/characterization and exposure evaluation [9]. Traditionally a chemical’s potential hazard(s) on human health are identified using standard animal (i.e., *in vivo*) toxicity testing of apical endpoints such as cancer [9–11]. In some cases, a serious hazard of a chemical such as it being CMR can result in restrictions irrespective of exposure level and use [10,12]. In most cases, however, the hazard characterization and subsequent risk assessment and classification and labeling of chemicals is more complex [7,12].

2.3.1 A Paradigm Shift in Toxicology

For the majority of the man-made chemicals none or only limited toxicity data are available [13,14], and use of classical regulatory toxicology *in vivo* tests to fill the large data gaps of the many thousands chemicals queued for risk assessment is practically impossible due to time and economic limitations [10,13,15–21]. Also, the ethically problematic animal toxicity studies do not always translate well to humans [22,23] and provide limited information on the actual mechanism(s) underlying the adverse outcome(s) [10,16,17,24,25]. To meet these challenges, regulatory toxicology has called for a paradigm shift to identify, develop and apply more sustainable and practical testing and non-testing methods that ultimately can replace animal testing [16,17,26–28]. Facing the challenge, the U.S. EPA together with the National Toxicology Program (NTP) asked the National Research Council (NRC) to develop a long-range vision and strategy for future toxicity testing, which resulted in the publication of the game-changing report from 2007 entitled ‘Toxicity Testing in the 21st Century: A Vision and a Strategy’ [16,17]. Here it is discussed how technological advances in molecular biology and computer science during the 20th and continuing into the 21st century can

help scientists identify cellular and molecular mechanisms in ‘toxicity pathways’ that may lead to adverse outcomes. The report envisions that understanding of chemical interaction with molecular mechanisms in ‘toxicity pathways’ can be used to reliably predict toxicity in a cost- and time-efficient way while reducing animal use and suffering [16,17,28].

Today, ten years after the report was released, agencies, academia and industry are continuously taking new initiatives to meet the paradigm shift. For example, new test and non-test methods are developed or optimized. HTS *in vitro* assays use either cell-free systems or cell-lines, preferably of human origin, to identify chemical interaction with mechanisms in ‘toxicity pathways’ [29–31]. The rationale is that a battery of such HTS *in vitro* assays can be used as a tool to identify and prioritize chemicals that should progress to further, more resource-demanding toxicological evaluation [30,32]. However, testing new sets of chemicals in medium- or high-throughput *in vitro* assays can also be costly and time-consuming due to the many ‘toxicity pathways’ molecular mechanisms that need to be covered and testing at multiple concentrations [33,34]. In addition, for some of these *in vitro* assays, use of animals is a necessity to get hold of the cell cultures [35]. Development and use of non-test methods, such as QSAR, to screen and prioritize chemicals for further testing can serve as a pre-filter for HTS testing [20,21,36] or be applied directly or indirectly (i.e., in groupings/read across methodology) to fill data gaps [7]. The alternative methods, both *in vitro* and *in silico*, have already resulted in an ocean of data and lead to questions on how to best handle, assess and recognize the limitations of this data [37]. Linking mechanistic data from alternative methods to adverse outcomes at the organism or population level is another challenge being faced [37]. The regulatory system has not fully adapted to the use of mechanistic data from alternative methods but still mainly relies on animal toxicity data. Furthermore, regulators, who have been trained to make decisions based on apical endpoint data from animal studies, may be unfamiliar with and uncertain about the interpretation of this new type of data, which further limits its potential use in chemical risk assessment.

As chemical risk assessments combines knowledge on the hazardous potential of the given chemical with its level of exposure and use, another major challenge in risk assessment is to estimate the human exposure levels of the many thousands chemicals in our surroundings [13,38–42]. Also, current chemical risk assessment is based on the exposure and hazards associated with a single chemical but humans and wildlife are exposed to complex mixtures of natural and man-made chemicals, which may act through multiple ‘toxicity pathways’ and can cause additive or synergistic toxicity effects [43,44]. Parallel to the challenge of filling data gaps on toxicity and exposure levels for individual chemical substances, is the challenge of how to test and risk assess chemical mixtures

[45]. The exposure and mixture effect challenges and some suggested methods to address these are discussed elsewhere [43,45–47] and will not be further elaborated in this thesis.

2.3.2 ToxCast and Tox21 Programs

To face the challenge of filling the toxicity data gaps for the many thousands of man-made chemicals, the U.S. EPA National Center for Computational Toxicology (NCCT) launched the Toxicity Forecaster research program, known as ToxCast¹¹, in 2007 with the overall aim to “*use in vitro HTS approaches to support the development of improved toxicity prediction models*” (cit. from [37]) [24,37,48,49]. ToxCast is the U.S. EPA contribution to the Toxicity in the 21st Century (Tox21) program, which was initiated in 2008 as a U.S. federal ‘multiagency’ collaboration among the U.S. EPA, the Food and Drug Administration (FDA) and National Institutes of Health (NIH), including the National Center for Advancing Translational Sciences (NCATS) and the NTP at the National Institute of Environmental Health Sciences (NIEHS) [50,51]. Tox21 was a response to the NRC report ‘*Toxicity Testing in the 21st Century: A Vision and a Strategy*’ [16,17,27], which calls for a collaborative effort across the toxicology community to rely less on animal studies and more on *in vitro* tests using human cells and cellular components to identify chemicals with toxic effects. Although ToxCast and Tox21 share the same overall aims [37,48,52,53], they apply different approaches. In Tox21 the focus is on testing a large chemical inventory of around 10,000 substances (the full Tox21 set, 8,193 unique chemicals) in a small selection of HTS assays each year [24,53], while in ToxCast an EPA selected subset of the Tox21 chemicals, currently 3,726, are tested in many hundreds of assays to cover multiple ‘toxicity pathways’ [37,51,54].

The ToxCast chemical library consists of structurally diverse man-made compounds such as plasticizers, pesticides, phthalates, antimicrobials and food additives as well as approved and failed drugs [24,25,37]. The ToxCast program is being conducted in multiple phases. Phase I was completed in 2009 as a ‘Proof of concept’. In this phase 310 unique chemicals, mainly pesticides with accompanying animal toxicity data, were screened for approximately 700 HTS assay endpoints [24,37,49]. Next, ToxCast Phase II was initiated and includes 293 reproducible Phase I chemicals, a subset of 768 chemicals considered to have the highest priority of the EPA Tox21 set, as well as 799 unique chemicals, known as the ‘Endocrine 1000’ or E1K set [37]. The Phase II chemicals are screened for around 900 assay endpoints, including most of the original approximately 700 endpoints from Phase I, with the exception of the E1K set, which is screened only in a limited subset of Phase II endocrine-related assays [37]. In late 2014, ToxCast Phase III was started with new

¹¹ <https://www.epa.gov/chemical-research/toxicity-forecasting>

technologies and endpoints added, as well as including a new set of ~1900 unique EPA selected Tox21 chemicals of regulatory concern [36,37].

The inclusion criteria and procurement of EPA's Tox21 subset inventory, which currently consists of 3,726 chemicals, are described in [37]. All chemicals have undergone thorough quality reviews [55], and the chemical structures have undergone a standardized and validated procedure to produce EPA 'QSAR-ready' structures [37,56]. The assays in ToxCast and Tox21 are a compilation of biochemical assays, cell-based assays, complex culture assays and small animal models, and most of these were originally applied by the pharmaceutical industry [24,50]. The NIH NCATS high-throughput robotic screening system is used on some of the commercial assays [37]. A subset of the assays has been developed by U.S. EPA or NIH scientists as part of the ToxCast/Tox21 program [24,35,57,58]. Most assays were run in medium or high-throughput concentration-response for all chemicals [24], and in general assay data are considered to be of high quality and reproducibility [35,50,53,57]. In some cases, a tiered screening approach is applied, where the chemicals are first tested at a single high concentration, and chemicals exceeding a defined endpoint activity threshold are prioritized for concentration-response testing [57,58]. The raw concentration-response ToxCast data from different sources are processed through a U.S. EPA customized data analyzing pipeline in R¹², which results in a final 'hit-call' for each chemical-assay-endpoint [37,59]. A 'hit-call' of 1 (active) or 0 (inactive) for a chemical is based on a decision on whether a statistically significant concentration-response is modeled and takes into account outliers and general toxicity data such as cytotoxicity [59]. The full Tox21 dataset is processed through another but similar data analyzing pipeline by NIH [53,60,61].

Besides the Tox21/ToxCast programs other sources of HTS *in vitro* data exists. PubChem¹³ is a free, online database from NIH National Center for Biotechnology Information (NCBI) that provides structural information for millions of chemical structures and data on biological activities of small molecules [62]. Part of the ToxCast and Tox21 data are available from PubChem. Other similar online databases include ChEMBL¹⁴, BindingDB¹⁵, ChemProt¹⁶ and CTD¹⁷.

The data used in project 3.1 are from ToxCast, and project 3.4 data originate from ToxCast and Tox21 assays. The data in project 3.2 were from NIH NCATS but on another chemical collection called NCGC (NCATS Chemical Genomics Center), which consists primarily of drugs [63,64]. The models in projects 3.1, 3.2 and 3.4 were all developed in close collaboration with the data providers.

¹² <https://www.r-project.org/about.html>

¹³ <https://pubchem.ncbi.nlm.nih.gov/>

¹⁴ <https://www.ebi.ac.uk/chembl/db/>

¹⁵ <http://www.bindingdb.org/bind/index.jsp>

¹⁶ <http://potentia.cbs.dtu.dk/ChemProt/>

¹⁷ <http://ctdbase.org/>

In project 3.3, data were curated from the PubChem database [62]. More information on the data can be found in the respective project chapters.

2.3.3 Adverse Outcome Pathways

As mentioned earlier, the use of mechanistic data from alternative methods, such as the ToxCast and Tox21 HTS *in vitro* data, in a regulatory toxicology context has faced multiple challenges [11,24]. To meet the challenge of how to link mechanistic results from alternative methods to adverse effects at the organism or population level, OECD initiated the development of AOPs in 2012 [65]. The AOP framework is an expansion of NRCs 'toxicity pathways' [16,17] and the Mode-of-Action (MoA) concept (Figure 5) [66–68], and it aims to simplify complex biological systems by relating molecular mechanisms to adverse effects in a one-way scheme. Descriptions of biological pathways is not a new concept, but has been made by scientists for decades. The novelty in the AOP framework is to systematize, standardize and simplify the pathways to make them useful in a regulatory context.

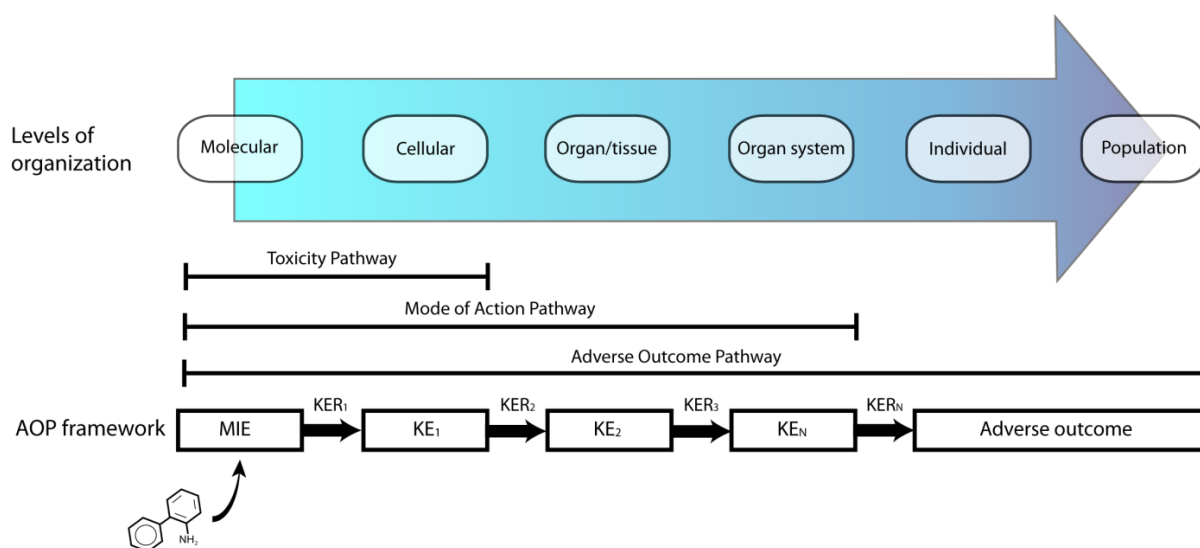


Figure 5. The AOP framework

An AOP endeavors to make a simple representation of existing knowledge concerning causal linkages between an MIE and a cascade of intermediate key events (KEs) at subcellular, cellular, tissue and/or organ levels that lead to a specific adverse outcome (AO) at individual or population level (Figure 5) [10,66,69]. An AO can be explained by multiple AOPs in a so-called AOP network [70], just as an MIE or a KE may be included in several AOPs with different AOs [11]. The AOP conceptual framework provides the biological context to alternative data with the objective to make e.g. regulators more familiar with and confident in the use of mechanistic data from alternative methods in e.g. WoE assessments or integrated testing strategies (ITS) for chemical risk assessment. Also,

well-constructed AOPs can help identify where existing testing or non-testing approaches can facilitate regulatory decision making, and drive development of new key *in vitro* assays and *in silico* models [10,11]. Furthermore, information from AOPs can be used in the design and refinement of *in vivo* experiments to get as much relevant information out of the animals used. The ultimate and long-term regulatory goal of the AOP framework is to replace animal toxicity testing of a chemical with alternative methods for effects on MIEs and/or KEs levels.

In 2014, OECD in collaboration with the U.S. EPA, the U.S. Army Engineer Research & Development Center (ERDC) and the European Commission (EC) Joint Research Center (JRC) launched the AOP Knowledgebase (KB)¹⁸. The AOP-KB integrates four individually developed platforms to more effectively allow stakeholders to develop, review and comment on AOPs. The AOP-Wiki¹⁹, developed by the U.S. EPA and EC JRC, is one of the platforms in the AOP-KB and serves as a central repository for all AOPs under development. The AOPs in the AOP-Wiki are dynamic and at different stages in their development. In addition, OECD with financial support from the EC have developed Effectopedia²⁰, an open-knowledge and structured online platform able to display quantitative information in AOPs.

2.3.4 Integrated Approaches to Testing and Assessment

In addition to the AOP initiative, OECD introduced the IATA concept [71] to assist in the paradigm shift within regulatory toxicology [67]. In IATA a defined question regarding a chemical's (or a group of chemicals) hazard identification, characterization or risk assessment is answered by taking a systematic and iterative approach to integrate existing information from multiple methodologies and techniques, including QSAR, read-across, toxicogenomics, *in vitro* and *in vivo*, with the identification of data gaps and a judicious generation of new data [10,67]. The main benefits expected from the use of IATAs include reduction, refinement and replacement of animal testing (i.e. the 3Rs), more cost-effective and efficient testing and assessment as well as the generation of more extensive and reliable data [67].

An IATA can range from the more flexible and less formalized judgement-based approaches to the more structured and rigid rule-based approaches that leaves little or no room for expert choices [10,67,72,73]. The choice of IATA depends on the specific decision-making and its context. Overall, existing and new data are continuously used in a WoE assessment to inform regulatory decisions and when an acceptable level of information is met, a final regulatory decision can be reached. The IATA decision procedure integrates gathered information on a chemical's exposure level/use, ADME

¹⁸ <http://aopkb.org>

¹⁹ <https://aopwiki.org>

²⁰ <https://effectopedia.org>

(absorption, distribution, metabolism and excretion) and toxicity in a WoE assessment approach to reach the decision on the endpoint of concern (Figure 6).

The AOP concept can be included in an IATA to provide the biological rationale in the decision making and to identify MIEs or KEs for which methods and data exist or for which new testing or non-testing methods are desirable [10,74]. If existing testing, e.g. HTS *in vitro* assays, or non-testing, e.g. QSARs, methods are available for an MIE/KE these can be used for generating new data to inform the IATA. In cases where *in vitro* assays or QSARs are missing/unavailable for an MIE/KE assessed to be relevant in the AOP-based IATA, the development of new testing and non-testing methods may be initiated (Figure 6).

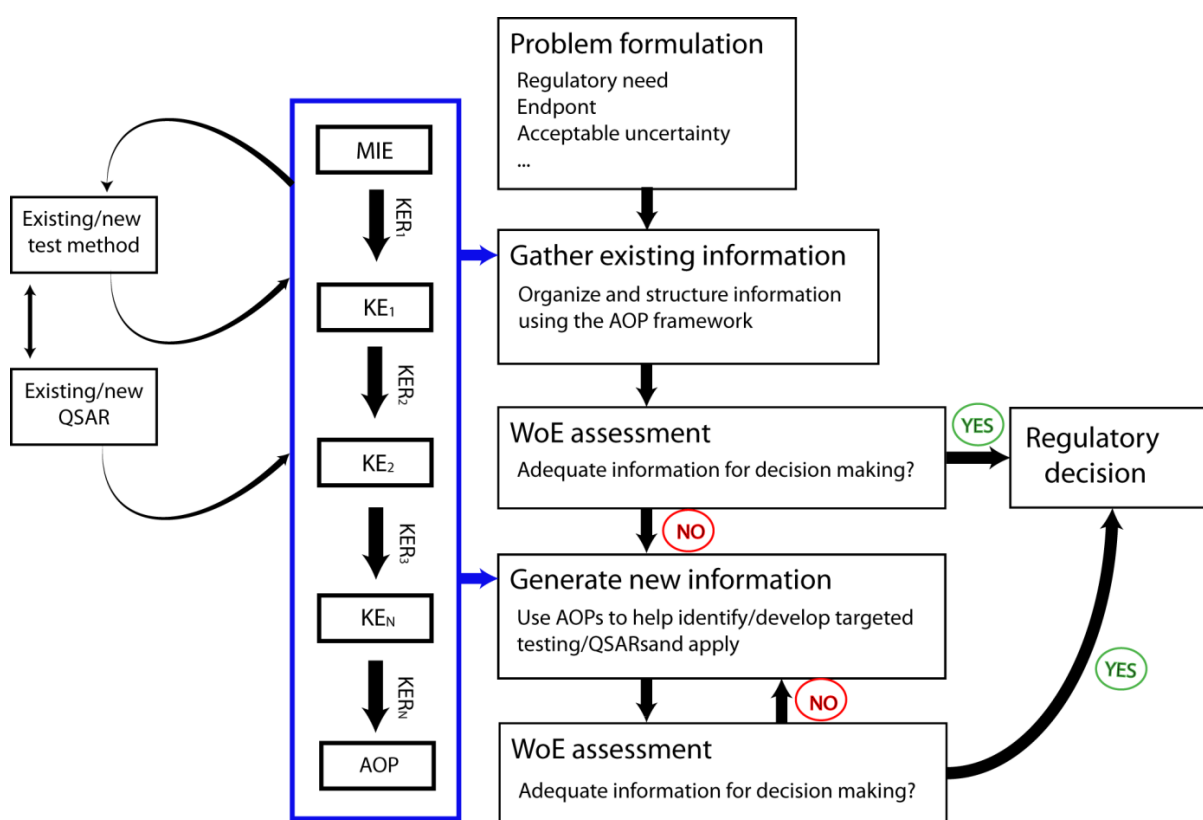


Figure 6. Illustration of an AOP-based IATA

2.3.5 Registration, Evaluation and Autorisation of Chemicals

The EU chemical legislation, REACH, was put into force in June 2007 [7,75] to ensure the safe use of chemicals with minimal risk for humans and the environment as well as to promote the development of alternatives to animal testing and enhance innovation and competitiveness in the industry [7]. One of the key principles in REACH is that the responsibility for demonstrating the safe use of chemicals lies with the industry/registrants [76]. Multiple deadlines for the registration of substances under REACH have been set since its implementation in 2007 with the final registration deadline in June 2018 for the lowest tonnage substances, i.e. less than 10 tonnes per year. The

registration deadlines have put pressure on the industry/registrant to collect the necessary toxicity data for the more than 70,000 anticipated registrations [19,77]. While applying a precautionary risk assessment approach, REACH is also cutting edge in the use of alternative testing and non-testing methods for regulatory purposes. In Articles 13 and 25 of REACH it is clearly stated that vertebrate testing should only be performed as a last resort after considering all other options such as gathering all existing information available on the substance, including information from alternative methods such as *in vitro* methods and (Q)SARs [7].

The minimum toxicity testing requirements for a registered substance under REACH depends on the quantity of the substance manufactured or imported into EU in tonnes per year, with higher requirements the higher the quantity [7]. The standard information requirements for the different tonnages are described in Annexes VII to X of REACH [7]. QSARs can potentially be used to meet standard information requirements at all tonnages levels if they are assessed adequate for the specific purpose. Overall, (Q)SAR results can be used instead of testing for regulatory purposes when the following conditions are met: 1) the results are derived from a scientifically valid (Q)SAR model following the OECD validation principles (see section 2.2.2), 2) the predicted substance falls within the QSAR model's AD, 3) the predictions are assessed to be adequate for the purpose of classification and labelling and/or risk assessment, and 4) adequate and reliable documentation on the applied model is provided [76]. These conditions are best documented in QMRF and QPRF. If some of the information elements in the conditions are missing or are inadequate, the (Q)SAR predictions may still be used in a WoE assessment approach in e.g. in an AOP-informed IATA [10,76]. At quantities of 10 or more tonnes per year the chemical substance has to be evaluated for reproductive toxicity according to the standard information requirements listed in Annex VIII to X [7]. In 2014, the extended one-generation reproductive toxicity study (EOGRTS) [78] replaced the two-generation reproductive study in column 1 of point 8.7.3 of Annexes IX and X [7,79] and was included in the EU test method regulation amendment [80]. DNT testing using e.g. cohort 2A/2B in EOGRTS is only required in REACH in case of serious concerns [7,18]. Triggers of such concerns are currently being identified in a close collaboration between the European Chemicals Agency (ECHA), member states and stakeholders and should result in a guidance document [81]. Suggestion for such triggers could be evidence from alternative methods on chemical interaction with MIEs or KEs in AOPs for DNT outcomes [10], for example some of the thyroid-related AOPs under development [10,82–86].

Endocrine disruption represents another potential gap in REACH requested dossier information (as well as other EU regulations) [18]. On June 15th 2016, the EC published a draft on its long-awaited and debated criteria for the identification of EDCs in a Communication together with an impact

assessment report setting out the criteria implications on regulations and their implementations [87,88]. The criteria have been criticized by politicians, scientists, NGOs and a number of member states, including Denmark, to be too weak to protect humans and the environment against adverse effects from EDCs [89,90].

References

- [1] J.A. Timbrell, *Principles of Biochemical Toxicology*, 4th ed., Informa Healthcare Inc., 2009.
- [2] CAS, CAS Assigns the 100 Millionth CAS Registry Number to a Substance Designed to Treat Acute Myeloid Leukemia, (2015). <http://www.cas.org/news/media-releases/100-millionth-substance> (accessed March 16, 2017).
- [3] US EPA, *Persistent Organic Pollutants: A Global Issue, A Global Response*, (2009). <https://www.epa.gov/international-cooperation/persistent-organic-pollutants-global-issue-global-response> (accessed March 16, 2017).
- [4] CCPSA, *Canada Consumer Product Safety Act*, (2010). <http://laws-lois.justice.gc.ca/eng/acts/C-1.68/page-1.html#h-3> (accessed March 16, 2017).
- [5] CSCL, *Chemical Substances Control Law*, (2017). http://www.meti.go.jp/policy/chemical_management/english/cscl/ (accessed March 16, 2017).
- [6] EU, *Regulation (EU) No 528/2012 of the European Parliament and of the Council 22 May 2012 concerning the making available on the market and use of biocidal products*, (2012). <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012R0528&from=EN>.
- [7] REACH, *Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)*, (2006). <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02006R1907-20161011&from=EN>.
- [8] TSCA, *Summary of the Toxic Substances Control Act*, (2017). <https://www.epa.gov/laws-regulations/summary-toxic-substances-control-act> (accessed March 16, 2017).
- [9] L. Hopper, F. Oehme, Chemical risk assessment: a review, *Vet. Hum. Toxicol.* 31 (1989) 543–554. <http://europepmc.org/abstract/med/2694585>.
- [10] K.E. Tollefsen, S. Scholz, M.T. Cronin, S.W. Edwards, J. de Knecht, K. Crofton, N. Garcia-Reyero, T. Hartung, A. Worth, G. Patlewicz, Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA), *Regul. Toxicol. Pharmacol.* 70 (2014) 629–640. doi:10.1016/j.yrtph.2014.09.009.
- [11] C. Wittwehr, H. Aladjov, G. Ankley, H.J. Byrne, J. de Knecht, E. Heinzle, G. Klambauer, B. Landesmann, M. Luijten, C. MacKay, G. Maxwell, M.E. (Bette) Meek, A. Paini, E. Perkins, T. Sobanski, D. Villeneuve, K.M. Waters, M. Whelan, How Adverse Outcome Pathways Can Aid the Development and Use of Computational Prediction Models for Regulatory Toxicology, *Toxicol. Sci.* 155 (2017) 326–336. doi:10.1093/toxsci/kfw207.
- [12] CLP, *Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures*, (2008). <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008R1272&from=EN> (accessed March 16, 2017).
- [13] R. Judson, A. Richard, D.J. Dix, K. Houck, M. Martin, R. Kavlock, V. Dellarco, T. Henry, T. Holderman, P. Sayre, S. Tan, T. Carpenter, E. Smith, The Toxicity Data Landscape for Environmental Chemicals, *Environ. Health Perspect.* 117 (2009) 685–695. doi:10.1289/ehp.0800168.
- [14] T.G. Neltner, H.M. Alger, J.E. Leonard, M. V. Maffini, Data gaps in toxicity testing of chemicals allowed in food in the United States, *Reprod. Toxicol.* 42 (2013) 85–94. doi:10.1016/j.reprotox.2013.07.023.

- [15] T. Hartung, Toxicology for the twenty-first century, *Nature*. 460 (2009) 208–212. doi:10.1038/460208a.
- [16] NRC, Toxicity Testing in the 21st Century: A Vision and a Strategy (Report in brief), 2007. http://dels.nas.edu/resources/static-assets/materials-based-on-reports/reports-in-brief/Toxicity_Testing_final.pdf (accessed December 20, 2016).
- [17] NRC, Toxicity Testing in the Twenty-first Century: A Vision and a Strategy, (2007). <http://dels.nas.edu/Report/Toxicity-Testing-Twenty-first/11970> (accessed March 13, 2017).
- [18] C. Rovida, How are reproductive toxicity and developmental toxicity addressed in REACH dossiers?, *ALTEX*. 28 (2011) 273–294. doi:10.14573/altex.2011.4.273.
- [19] C. Rovida, T. Hartung, Re-evaluation of animal numbers and costs for in vivo tests to accomplish REACH legislation requirements for chemicals - a report by the Transatlantic Think Tank for Toxicology (t4), *ALTEX*. 26 (2009) 187–208. doi:10.14573/altex.2009.3.187.
- [20] US-EPA NCCT, CERAPP -Collaborative Estrogen Receptor Activity Prediction Project, (2016). <https://www.epa.gov/chemical-research/cerapp-collaborative-estrogen-receptor-activity-prediction-project-0> (accessed March 16, 2017).
- [21] C.E. Willett, P.L. Bishop, K.M. Sullivan, Application of an Integrated Testing Strategy to the U.S. EPA Endocrine Disruptor Screening Program, *Toxicol. Sci.* 123 (2011) 15–25. doi:10.1093/toxsci/kfr145.
- [22] D. Fourches, J.C. Barnes, N.C. Day, P. Bradley, J.Z. Reed, A. Tropsha, Cheminformatics Analysis of Assertions Mined from Literature That Describe Drug-Induced Liver Injury in Different Species, *Chem. Res. Toxicol.* 23 (2010) 171–183. doi:10.1021/tx900326k.
- [23] C.A. LaLone, D.L. Villeneuve, D. Lyons, H.W. Helgen, S.L. Robinson, J.A. Swintek, T.W. Saari, G.T. Ankley, Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS): A Web-Based Tool for Addressing the Challenges of Cross-Species Extrapolation of Chemical Toxicity., *Toxicol. Sci.* 153 (2016) 228–245. doi:10.1093/toxsci/kfw119.
- [24] R. Kavlock, K. Chandler, K. Houck, S. Hunter, R. Judson, N. Kleinstreuer, T. Knudsen, M. Martin, S. Padilla, D. Reif, A. Richard, D. Rotroff, N. Sipes, D. Dix, Update on EPA's ToxCast Program: Providing High Throughput Decision Support Tools for Chemical Risk Management, *Chem. Res. Toxicol.* 25 (2012) 1287–1302. doi:10.1021/tx3000939.
- [25] F. Shah, N. Greene, Analysis of Pfizer Compounds in EPA's ToxCast Chemicals-Assay Space, *Chem. Res. Toxicol.* 27 (2014) 86–98. doi:10.1021/tx400343t.
- [26] P. Anastas, K. Teichman, E.C. Hubal, Ensuring the safety of chemicals, *J. Expo. Sci. Environ. Epidemiol.* 20 (2010) 395–396. doi:10.1038/jes.2010.28.
- [27] D. Krewski, D. Acosta, M. Andersen, H. Anderson, J.C. Bailar, K. Boekelheide, R. Brent, G. Charnley, V.G. Cheung, S. Green, K.T. Kelsey, N.I. Kerkvliet, A.A. Li, L. McCray, O. Meyer, R.D. Patterson, W. Pennie, R.A. Scala, G.M. Solomon, M. Stephens, J. Yager, L. Zeise, Staff of Committee on Toxicity Test, Toxicity Testing in the 21st Century: A Vision and a Strategy, *J. Toxicol. Environ. Heal. Part B.* 13 (2010) 51–138. doi:10.1080/10937404.2010.483176.
- [28] NTP, Toxicology in the 21 st Century: The Role of the National Toxicology Program, (2004). https://ntp.niehs.nih.gov/ntp/main_pages/ntpvision.pdf (accessed March 16, 2017).
- [29] R. Judson, K. Houck, M. Martin, T. Knudsen, R.S. Thomas, N. Sipes, I. Shah, J. Wambaugh, K. Crofton, In Vitro and Modelling Approaches to Risk Assessment from the U.S. Environmental Protection Agency ToxCast Programme, *Basic Clin. Pharmacol. Toxicol.* 115 (2014) 69–76. doi:10.1111/bcpt.12239.

- [30] R.J. Kavlock, D.J. Dix, K.A. Houck, R.S. Judson, M.T. Martin, A.M. Richard, ToxCast: Developing predictive signatures for chemical toxicity, *AATEX*. 14 (2008) 623–627. http://www.asas.or.jp/jsaae/jsaae/zasshi/WC6_PC/paper623.pdf.
- [31] N.C. Kleinstreuer, J. Yang, E.L. Berg, T.B. Knudsen, A.M. Richard, M.T. Martin, D.M. Reif, R.S. Judson, M. Polokoff, D.J. Dix, R.J. Kavlock, K.A. Houck, Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms, *Nat. Biotechnol.* 32 (2014) 583–591. doi:10.1038/nbt.2914.
- [32] R. Judson, R. Kavlock, M. Martin, D. Reif, K. Houck, T. Knudsen, A. Richard, R.R. Tice, M. Whelan, M. Xia, R. Huang, C. Austin, G. Daston, T. Hartung, J.R. Fowle III, W. Wooge, W. Tong, D. Dix, Perspectives on validation of high-throughput assays supporting 21st century toxicity testing, *ALTEX*. 30 (2013) 51–56. doi:10.14573/altex.2013.1.051.
- [33] S.J. Capuzzi, R. Politi, O. Isayev, S. Farag, A. Tropsha, QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays, *Front. Environ. Sci.* 4 (2016) 1–7. doi:10.3389/fenvs.2016.00003.
- [34] Danishuddin, A.U. Khan, Descriptors and their selection methods in QSAR analysis: paradigm for drug design, *Drug Discov. Today*. 21 (2016) 1291–1302. doi:10.1016/j.drudis.2016.06.013.
- [35] K.B. Paul, J.M. Hedge, D.M. Rotroff, M.W. Hornung, K.M. Crofton, S.O. Simmons, Development of a Thyroperoxidase Inhibition Assay for High-Throughput Screening, *Chem. Res. Toxicol.* 27 (2014) 387–399. doi:10.1021/tx400310w.
- [36] EDSP21 Work Plan, The Incorporation of In Silico Models and In Vitro High Throughput Assays in the Endocrine Disruptor Screening Program (EDSP) for Prioritization and Screening, (2011). https://www.epa.gov/sites/production/files/2015-07/documents/edsp21_work_plan_summary_overview_final.pdf (accessed March 13, 2017).
- [37] A.M. Richard, R.S. Judson, K.A. Houck, C.M. Grulke, P. Volarath, I. Thillainadarajah, C. Yang, J. Rathman, M.T. Martin, J.F. Wambaugh, T.B. Knudsen, J. Kancharla, K. Mansouri, G. Patlewicz, A.J. Williams, S.B. Little, K.M. Crofton, R.S. Thomas, ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology, *Chem. Res. Toxicol.* 29 (2016) 1225–1251. doi:10.1021/acs.chemrestox.6b00135.
- [38] K.L. Dionisio, A.M. Frame, M.-R. Goldsmith, J.F. Wambaugh, A. Liddell, T. Cathey, D. Smith, J. Vail, A.S. Ernstoff, P. Fantke, O. Jolliet, R.S. Judson, Exploring consumer exposure pathways and patterns of use for chemicals in the environment, *Toxicol. Reports*. 2 (2015) 228–237. doi:10.1016/j.toxrep.2014.12.009.
- [39] P.P. Egeghy, L.S. Sheldon, K.K. Isaacs, H. ??zkaynak, M.R. Goldsmith, J.F. Wambaugh, R.S. Judson, T.J. Buckley, Computational exposure science: An emerging discipline to support 21st-century risk assessment, *Environ. Health Perspect.* 124 (2016) 697–702. doi:10.1289/ehp.1509748.
- [40] P.P. Egeghy, R. Judson, S. Gangwal, S. Mosher, D. Smith, J. Vail, E.A. Cohen Hubal, The exposure data landscape for manufactured chemicals, *Sci. Total Environ.* 414 (2012) 159–166. doi:10.1016/j.scitotenv.2011.10.046.
- [41] M. Fryer, C.D. Collins, H. Ferrier, R.N. Colville, M.J. Nieuwenhuijsen, Human exposure modelling for chemical risk assessment: a review of current approaches and research and policy implications, *Environ. Sci. Policy*. 9 (2006) 261–274. doi:10.1016/j.envsci.2005.11.011.
- [42] J.F. Wambaugh, A. Wang, K.L. Dionisio, A. Frame, P. Egeghy, R. Judson, R.W. Setzer, High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals, *Environ. Sci. Technol.* 48 (2014) 12760–12767. doi:10.1021/es503583j.

- [43] A. Kortenkamp, T. Backhaus, M. Faust, State of the Art Report on Mixture Toxicity, (2009). http://ec.europa.eu/environment/chemicals/effects/pdf/report_mixture_toxicity.pdf (accessed March 14, 2017).
- [44] S.H. Safe, Hazard and Risk Assessment of Chemical Mixtures Using the Toxic Equivalency Factor Approach, *Environ. Health Perspect.* 106 (1998) 1051–1058. doi:10.2307/3434151.
- [45] A. Kienzler, S.K. Bopp, S. van der Linden, E. Berggren, A. Worth, Regulatory assessment of chemical mixtures: Requirements, current approaches and future perspectives, *Regul. Toxicol. Pharmacol.* 80 (2016) 321–334. doi:10.1016/j.yrtph.2016.05.020.
- [46] A. Kortenkamp, Ten Years of Mixing Cocktails: A Review of Combination Effects of Endocrine-Disrupting Chemicals, *Environ. Health Perspect.* 115 (2007) 98–105. doi:10.1289/ehp.9357.
- [47] A. Kortenkamp, Low dose mixture effects of endocrine disrupters and their implications for regulatory thresholds in chemical risk assessment, *Curr. Opin. Pharmacol.* 19 (2014) 105–111. doi:10.1016/j.coph.2014.08.006.
- [48] D.J. Dix, K.A. Houck, M.T. Martin, A.M. Richard, R.W. Setzer, R.J. Kavlock, The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals, *Toxicol. Sci.* 95 (2007) 5–12. doi:10.1093/toxsci/kfl103.
- [49] R.S. Judson, K.A. Houck, R.J. Kavlock, T.B. Knudsen, M.T. Martin, H.M. Mortensen, D.M. Reif, D.M. Rotroff, I. Shah, A.M. Richard, D.J. Dix, In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project, *Environ. Health Perspect.* 118 (2010) 485–492. doi:10.1289/ehp.0901392.
- [50] F.S. Collins, G.M. Gray, J.R. Bucher, TOXICOLOGY: Transforming Environmental Health Protection, *Science* 319 (2008) 906–907. doi:10.1126/science.1154619.
- [51] R.J. Kavlock, C.P. Austin, R.R. Tice, Toxicity Testing in the 21st Century: Implications for Human Health Risk Assessment, *Risk Anal.* 29 (2009) 485–487. doi:10.1111/j.1539-6924.2008.01168.x.
- [52] A. Abdelaziz, H. Spahn-Langguth, K.-W. Schramm, I. V. Tetko, Consensus Modeling for HTS Assays Using In silico Descriptors Calculates the Best Balanced Accuracy in Tox21 Challenge, *Front. Environ. Sci.* 4 (2016) 1–12. doi:10.3389/fenvs.2016.00002.
- [53] R. Huang, M. Xia, S. Sakamuru, J. Zhao, S.A. Shahane, M. Attene-Ramos, T. Zhao, C.P. Austin, A. Simeonov, Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization, *Nat. Commun.* 7 (2016) 10425. doi:10.1038/ncomms10425.
- [54] R.R. Tice, C.P. Austin, R.J. Kavlock, J.R. Bucher, Improving the Human Hazard Characterization of Chemicals: A Tox21 Update, *Environ. Health Perspect.* 121 (2013) 756–765. doi:10.1289/ehp.1205784.
- [55] U.S. EPA, ToxCast Chemical Inventory: Data Management and Data Quality Considerations, 2014. https://www.epa.gov/sites/production/files/2015-08/documents/toxcast_chemicals_qa_qc_management_141204.pdf (accessed January 13, 2017).
- [56] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A.M. Richard, C.M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E.B. Wedebye, F. Grisoni, G.F. Mangiatordi, G.M. Incisivo, H. Hong, H.W. Ng, I. V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N.G. Nikolov, O. Nicolotti, P.L. Andersson, Q. Zang, R. Politi, R.D. Beger, R. Todeschini, R. Huang, S. Farag, S.A. Rosenberg, S. Slavov, X. Hu, R.S. Judson, CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ. Health Perspect.* 124 (2016) 1023–1033. doi:10.1289/ehp.1510267.

- [57] A.L. Karmaus, C.M. Toole, D.L. Filer, K.C. Lewis, M.T. Martin, High-Throughput Screening of Chemical Effects on Steroidogenesis Using H295R Human Adrenocortical Carcinoma Cells, *Toxicol. Sci.* 150 (2016) 323–332. doi:10.1093/toxsci/kfw002.
- [58] K. Paul Friedman, E.D. Watt, M.W. Hornung, J.M. Hedge, R.S. Judson, K.M. Crofton, K.A. Houck, S.O. Simmons, Tiered High-Throughput Screening Approach to Identify Thyroperoxidase Inhibitors Within the ToxCast Phase I and II Chemical Libraries, *Toxicol. Sci.* 151 (2016) 160–180. doi:10.1093/toxsci/kfw034.
- [59] D.L. Filer, P. Kothiya, W.R. Setzer, R.S. Judson, M.T. Martin, The ToxCast™ Analysis Pipeline: An R Package for Processing and Modeling Chemical Screening Data, 2015. https://www.epa.gov/sites/production/files/2015-08/documents/pipeline_overview.pdf (accessed January 11, 2017).
- [60] R. Huang, M. Xia, M.-H. Cho, S. Sakamuru, P. Shinn, K.A. Houck, D.J. Dix, R.S. Judson, K.L. Witt, R.J. Kavlock, R.R. Tice, C.P. Austin, Chemical Genomics Profiling of Environmental Chemical Modulation of Human Nuclear Receptors, *Environ. Health Perspect.* 119 (2011) 1142–1148. doi:10.1289/ehp.1002952.
- [61] J. Inglese, D.S. Auld, A. Jadhav, R.L. Johnson, A. Simeonov, A. Yasgar, W. Zheng, C.P. Austin, Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries, *Proc. Natl. Acad. Sci.* 103 (2006) 11473–11478. doi:10.1073/pnas.0604348103.
- [62] Y. Wang, T. Suzek, J. Zhang, J. Wang, S. He, T. Cheng, B.A. Shoemaker, A. Gindulyte, S.H. Bryant, PubChem BioAssay: 2014 update, *Nucleic Acids Res.* 42 (2014) D1075–D1082. doi:10.1093/nar/gkt978.
- [63] R. Huang, N. Southall, Y. Wang, A. Yasgar, P. Shinn, A. Jadhav, D.-T. Nguyen, C.P. Austin, The NCGC Pharmaceutical Collection: A Comprehensive Resource of Clinically Approved Drugs Enabling Repurposing and Chemical Genomics, *Sci. Transl. Med.* 3 (2011). doi:10.1126/scitranslmed.3001862.
- [64] S.J. Shukla, S. Sakamuru, R. Huang, T.A. Moeller, P. Shinn, D. VanLeer, D.S. Auld, C.P. Austin, M. Xia, Identification of Clinically Used Drugs That Activate Pregnane X Receptors, *Drug Metab. Dispos.* 39 (2011) 151–159. doi:10.1124/dmd.110.035105.
- [65] OECD, Proposal for a template, and guidance on developing and assessing the completeness of adverse outcome pathways, (2012). <http://www.oecd.org/chemicalsafety/testing/49963554.pdf> (accessed March 16, 2017).
- [66] G.T. Ankley, R.S. Bennett, R.J. Erickson, D.J. Hoff, M.W. Hornung, R.D. Johnson, D.R. Mount, J.W. Nichols, C.L. Russom, P.K. Schmieder, J.A. Serrano, J.E. Tietge, D.L. Villeneuve, Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment, *Environ. Toxicol. Chem.* 29 (2010) 730–741. doi:10.1002/etc.34.
- [67] A.P. Worth, G. Patlewicz, Integrated Approaches to Testing and Assessment, in: Chantra Eskes, Maurice Whelan (Eds.), *Valid. Altern. Methods Toxic. Test.*, Springer International Publishing, 2016: pp. 317–342. doi:10.1007/978-3-319-33826-2_13.
- [68] R.T. Zoeller, K.M. Crofton, Mode of Action: Developmental Thyroid Hormone Insufficiency—Neurological Abnormalities Resulting From Exposure to Propylthiouracil, *Crit. Rev. Toxicol.* 35 (2005) 771–781. doi:10.1080/10408440591007313.
- [69] N.C. Kleinstreuer, K. Sullivan, D. Allen, S. Edwards, D.L. Mendrick, M. Embry, J. Matheson, J.C. Rowlands, S. Munn, E. Maull, W. Casey, Adverse outcome pathways: From research to regulation scientific workshop report, *Regul. Toxicol. Pharmacol.* 76 (2016) 39–50. doi:10.1016/j.yrtph.2016.01.007.

- [70] D. Knapen, L. Vergauwen, D.L. Villeneuve, G.T. Ankley, The potential of AOP networks for reproductive and developmental toxicity assay development, *Reprod. Toxicol.* 56 (2015) 52–55. doi:10.1016/j.reprotox.2015.04.003.
- [71] OECD, Workshop on Integrated Approaches to Testing and Assessment, 2008. [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2008\)10&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2008)10&doclanguage=en) (accessed January 13, 2017).
- [72] T. Hartung, T. Luechtefeld, A. Maertens, A. Kleensang, Food for Thought ... Integrated Testing Strategies for Safety Assessments, *ALTEX*. 30 (2013) 3–18. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3800026/pdf/nihms516171.pdf> (accessed March 16, 2017).
- [73] OECD, New guidance document on an Integrated Approach on Testing and Assessment (IATA) for skin corrosion and irritation, (2014). [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2014\)19&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2014)19&doclanguage=en) (accessed March 16, 2017).
- [74] G. Patlewicz, C. Kuseva, A. Kesova, I. Popova, T. Zhechev, T. Pavlov, D.W. Roberts, O. Mekenyan, Towards AOP application – Implementation of an integrated approach to testing and assessment (IATA) into a pipeline tool for skin sensitization, *Regul. Toxicol. Pharmacol.* 69 (2014) 529–545. doi:10.1016/j.yrtph.2014.06.001.
- [75] ECHA, Understanding REACH, (2017). <https://echa.europa.eu/regulations/reach/understanding-reach> (accessed March 16, 2017).
- [76] ECHA, Guidance on information requirements and chemical safety assessment - Chapter R.6: QSARs and grouping of chemicals, (2008). https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf (accessed March 16, 2017).
- [77] ChemistryViews, Get fit for the 2018 REACH Registration deadline, *ChemViews*. (2015). doi:10.1002/chemv.201200029.
- [78] OECD TG 443, Test No. 443: Extended One-Generation Reproductive Toxicity Study, OECD Publishing, 2012. doi:10.1787/9789264185371-en.
- [79] OECD TG 416, Test No. 416: Two-Generation Reproduction Toxicity, OECD Publishing, 2001. doi:10.1787/9789264070868-en.
- [80] EC, Commission Regulation (EU) No 900/2014 of 15 July 2014 amending, for the purpose of its adaptation to technical progress, Regulation (EC) No 440/2008 laying down test methods pursuant to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), (2014). <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0900>.
- [81] EC, Commission Regulation (EU) 2015/282 of 20 February 2015 amending Annexes VIII, IX and X to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as regards the Extended One-Generation Reproductive Toxicity Study, (2015). <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R0282&rid=1>.
- [82] AOP-134, Sodium Iodide Symporter (NIS) Inhibition and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/134> (accessed March 13, 2017).
- [83] AOP-152, Interference with thyroid serum binding protein transthyretin and subsequent adverse human neurodevelopmental toxicity, (2017). <https://aopwiki.org/aops/152> (accessed March 13, 2017).

- [84] AOP-42, Inhibition of Thyroperoxidase and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/42> (accessed March 13, 2017).
- [85] AOP-54, Inhibition of Na⁺/I⁻ symporter (NIS) decreases TH synthesis leading to learning and memory deficits in children, (2017). <https://aopwiki.org/aops/54> (accessed March 13, 2017).
- [86] AOP-8, Upregulation of Thyroid Hormone Catabolism via Activation of Hepatic Nuclear Receptors, and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/8> (accessed March 13, 2017).
- [87] EC, Communication from the Commission to the European Parliament and the Council: on endocrine disruptors and the draft Commission acts setting out scientific criteria for their determination in the context of the EU legislation on plant protection products and, (2016). http://ec.europa.eu/health/sites/health/files/endocrine_disruptors/docs/com_2016_350_en.pdf (accessed March 16, 2017).
- [88] EC, Commission staff working document impact assessment: Defining criteria for identifying endocrine disruptors in the context of the implementation of the plant protection products regulation and biocidal products regulation, (2016). http://ec.europa.eu/health/sites/health/files/endocrine_disruptors/docs/2016_impact_assessment_en.pdf (accessed March 16, 2017).
- [89] Altinget.dk, Forbrugerråd: 100 forskere tager ikke fejl om hormonforstyrrende stoffer - Altinget: miljø, (2017). <http://www.altinget.dk/miljoe/artikel/forbrugerraad-100-forskere-tager-ikke-fejl-om-hormonforstyrrende-stoffer> (accessed March 16, 2017).
- [90] BfR, International Expert Meeting on Endocrine Disruptors - BfR, (2016). http://www.bfr.bund.de/en/international_expert_meeting_on_endocrine_disruptors-197246.html (accessed March 16, 2017).

PART III - Projects

3.1 QSAR Models for TPO Inhibition *In Vitro*

3.1.1 Manuscript in Preparation

QSAR Models for Thyroperoxidase Inhibition and Screening of U.S. and EU Chemical Inventories

Rosenberg, S.A.^a, Watt, E.D.^{b,c}, Judson, R.S.^b, Simmons, S.O.^b, Paul Friedman, K.^b, Dybdahl, M.^{a1},
Nikolov, N.G.^{a1}, and Wedebye, E.B.^{a1*}

- a. *Division of Diet, Disease Prevention and Toxicology, National Food Institute, Technical University of Denmark, Kemitorvet, Building 202, 2800 Kgs. Lyngby, Denmark*
- b. *National Center for Computational Toxicology, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA*
- c. *Current Address: Computational ADME Group, Department of Pharmacokinetics, Dynamics, and Metabolism, Pfizer Worldwide Research & Development, Groton, CT 06340, USA*

¹*Contributed equally*

**Corresponding author: ebawe@food.dtu.dk, +45 35887604*

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views of policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Abstract

Thyroperoxidase (TPO) is the enzyme that synthesizes thyroid hormones (THs). TPO inhibition by chemicals can result in decreased TH levels and developmental neurotoxicity, and therefore identification of TPO inhibition is of high relevance in safety evaluation of chemicals. In the present study, we developed two global quantitative structure-activity relationship (QSAR) models for TPO inhibition *in vitro*. Rigorous cross- and blinded external validations demonstrated that the first model, QSAR1, built from a training set of 877 ToxCast chemicals, was robust and highly predictive with balanced accuracies of 80.6% (SD = 4.6%) and 85.3%, respectively. The external validation test set was subsequently merged with the training set to constitute a larger training set totaling 1,519 ToxCast chemicals for a second model, QSAR2, which underwent robust cross-validation with a balanced accuracy of 82.7% (SD = 2.2%). An analysis of QSAR2 identified the ten most discriminating structural features for TPO inhibition and non-inhibition, respectively. Both models were used to screen 72,524 REACH substances and 32,197 U.S. EPA substances, and QSAR2 with the expanded training set had approximately 10% larger coverages compared to QSAR1. Of the substances predicted within QSAR2's applicability domain, 8,790 (19.3%) REACH substances and 7,166 (19.0%) U.S. EPA substances, respectively, were predicted to be TPO inhibitors. A case study on butyl hydroxyanisole (BHA), which is used as an antioxidant, was included to exemplify how predictions from the developed QSAR2 model may aid in elucidating the modes of action in adverse outcomes of chemicals. Overall, predictions from QSAR2 can for example be used in priority setting of chemicals and in read-across cases or weight-of-evidence assessments.

Keywords

QSAR; thyroperoxidase (TPO) inhibition; Adverse Outcome Pathway (AOP); Screening; BHA; REACH;

Abbreviations

AD, applicability domain; AOP, adverse outcome pathway; AUR, Amplex®UltraRed; DIT, diiodotyrosine; BHA, butylated hydroxyanisole; DNT, developmental neurotoxicity; DTU, Technical University of Denmark; EDSP, Endocrine Disruption Screening Program; EPA, Environmental Protection Agency; Food, National Food Institute; HPT, hypothalamus-pituitary-thyroid; HTS, high-throughput screening; IATA, integrated approaches to testing and assessment; KE, key event; LPDM, Leadscape® Predictive Data Miner; MIE, molecular initiating event; MIT, monoiodotyrosine; NCCT, National Center for Computational Toxicology; OECD, Organisation for Economic Co-operation and Development; PLR, partial logistic regression; PRS, pre-registered substances; QSAR, quantitative structure-activity relationship; rT3, reverse triiodothyronine; SD, standard deviation; T3, triiodothyronine; T4, thyroxine; Tg, thyroglobulin; TH, thyroid hormone; TPO, thyroperoxidase; TR, thyroid hormone receptors; TRE, thyroid hormone response elements; WoE, weight-of-evidence

1. Introduction

Thyroid hormones (THs) participate in multiple biological processes from early development and throughout adulthood [1–3]. In the fetus and neonate, THs play an essential role in neurodevelopment [4], where they are involved in neuron differentiation, proliferation and migration, dendritic branching and synaptogenesis, and myelination [1,5]. In early gestation, the fetus depends entirely on maternally-derived THs until the fetal thyroid gland becomes functional at approximately gestational week 12 in humans and gestational day 17-18 in rats [1,6,7]. Maternal THs continue to contribute to fetal TH levels throughout gestation in both humans and rats [1,6]. Studies have shown that even a moderate and transient decrease in maternal TH levels during pregnancy is associated with permanent adverse neurological changes in the offspring [8]. In animal models and humans altered cognition, socialization, and motor function as well as hearing loss have been observed following moderate to severe hypothyroidism [6,9–17]. Even low levels of TH insufficiency during fetal development may result in measurable IQ deficits in children [9,13–18]. In adulthood, dysregulated TH levels can give reversible clinical symptoms of hypo- or hyperthyroidism [8] and are correlated with pathological processes involved in adverse outcomes such as cancer, obesity and type II diabetes mellitus [19,20].

Humans are exposed to tens of thousands of man-made chemicals through food, drugs, air, water and consumer products [21–24]. Large data gaps exist for most of these xenobiotics on their potential thyroid disrupting properties [25]. Xenobiotics can disturb TH homeostasis through many different mechanisms, including altered TH synthesis, transport, metabolism, and receptor activation as well as disruption of the HPT axis [10,25–28]. The same xenobiotic may act through more than one mechanism [25]. Because of the severity of the adverse effects that can be expected from chemical disruption of thyroid homeostasis, especially during early development, there is a need to develop a strategy for the identification and testing of thyroid-active compounds. As a step towards replacing expensive and time-consuming whole animal studies with alternative methods in chemical risk assessments, the Organisation for Economic Co-operation and Development (OECD) launched a new program on the development of Adverse Outcome Pathways (AOPs) in 2012 [29]. An AOP describes the sequential chain of causally linked events at different levels of biological organization starting from a so-called molecular initiating event (MIE) going through a number of downstream linked key events (KEs), and ends at an adverse health or ecotoxicological effect [29,30]. According to the OECD, AOPs are the central element of a toxicological knowledge framework to support chemical risk assessment based on mechanistic reasoning. AOPs can help industry and regulators use results from alternative methods, such as *in vitro* and *in silico* methods, in chemical risk assessments [31], e.g. by applying the AOP in OECDs Integrated Approaches to Testing Assessment (IATA) context

[29,32,33]. Multiple thyroid-related AOPs have been suggested [34,35]. One AOP under development determined to have a strong overall weight-of-evidence describes a series of linked events from the MIE, thyroperoxidase (TPO) inhibition, leading to hypothyroxinemia, and resulting in altered neurodevelopment and neurological dysfunction in the offspring [41, see also 4 and 19]. TPO is a heme-containing multifunction enzyme essential in TH synthesis [37,38]. Recently, a high-throughput screening (HTS) *in vitro* assay for TPO inhibition was developed by the U.S. Environmental Protection Agency (EPA) National Center for Computational Toxicology (NCCT) [39] and used to screen 1,126 ToxCast Phase I and II chemicals including structurally diverse environmental chemicals and failed drugs [34,40,41]. The assay is based on microsomes from rat thyroid tissue and requires the amount from approximately one rat to assess quantitative TPO inhibition of 1.5 chemicals [39]. An additional set of 771 ToxCast chemicals (known as the 'Endocrine 1000' or 'E1K' set) [41,42] was subsequently screened in the same HTS TPO inhibition assay (Simmons *et al.*, in prep).

The goal of the present study was to use the ToxCast data to develop *in silico* models, and apply the models to large inventories of man-made chemicals to predict their potential to inhibit TPO. For this purpose, we first used experimental TPO inhibition results for 1,126 ToxCast Phase I and II chemicals, including replicated samples, to prepare a training set of 877 unique chemicals, which was then used to train and cross-validate a global binary Quantitative Structure-Activity Relationship (QSAR) model. QSARs are mathematical models that relate chemical structure descriptors with an experimental continuous (e.g. EC₅₀) or categorical (e.g. positive/negative) activity. Once established, these *in silico* models can be used as a non-testing approach to predict the activities of untested chemical structures (an introduction to QSAR can e.g. be found in [43] and [44]). The E1K dataset was used to prepare a test set of 646 chemicals, which was applied to externally validate the QSAR model. Next, the test set was merged with the training set to form a larger training set of 1,519 unique chemicals, which was subsequently used for training and cross-validating a second QSAR model. An analysis of the structural features in the second QSAR model was performed to identify features that best discriminated TPO inhibitors from non-inhibitors. Both QSAR models were used to screen two large EU and U.S. chemical inventories containing man-made substances potentially present in e.g. the environment and consumer products for their possible TPO inhibition activity. The screened EU inventory consists of 72,524 REACH pre-registered substances (PRS) structures extracted from the online Danish (Q)SAR Database structure set [45,46]. Briefly, REACH pre-registration concerns existing substances that companies plan to register under REACH, the EUs chemicals regulation, as so-called phase-in substances. The U.S. inventory was originally curated by the U.S. EPA as a part of the CERAPP project [47] and contains 32,464 unique structures to which humans are potentially

exposed. The structures were curated from sources such as the ACToR CPCat database [21], the DSSTox database [48], the Canadian Domestic Substances List, the Endocrine Disruption Screening Program set and EPI Suite training and test sets [41,42,47]. Predictions from these screenings will inform a tiered approach to prioritize possible thyroid modulating chemicals for further evaluation and could be used, together with relevant AOP(s), in IATA weight-of evidence (WoE) risk assessments [29,33,49]. We also conducted a case study to highlight how the developed QSAR models for TPO inhibition can support hypotheses regarding the mode of action for chemical-induced adverse outcomes observed in *in vivo* studies.

2. Materials and Methods

2.1 Experimental Datasets

We used two datasets provided by U.S. EPA NCCT with chemical structure information and HTS screening results for TPO inhibition *in vitro* to train and validate two QSAR models. The chemicals screened contained diverse chemical structures including environmental and industrial chemicals as well as some failed drugs [41]. The chemicals in both datasets were not selected specifically for this project or based on suspected TPO inhibition activity, and the original datasets include internal replicated samples. The experimental results consisted of data from the HTS Amplex®UltraRed-thyroperoxidase (AUR-TPO) *in vitro* assay [39], which had further undergone a selectivity filtering procedure to identify potentially false positive results due to non-specific activity decrease in the AUR-TPO assay [34]. Briefly, all chemical structures were initially screened at a single, high concentration (~87.5µM). The chemicals associated with 20% or greater decreases in maximal TPO activity were subsequently screened for possible concentration-response. The concentration-response data were processed as described previously using the ToxCast data pipeline whereby each chemical was assigned a 'hit-call' of 1 if active in AUR-TPO, or a 'hit-call' of 0 if inactive in AUR-TPO [50]. Actives in the AUR-TPO assay were further processed through a selectivity filtering algorithm, which integrates results from cytotoxicity and luciferase inhibition assays to identify possible non-specific positive results in the AUR-TPO assay [34]. The chemical structures, assays, data analysis and selectivity filtering procedure have been described in more details previously [34,39,40,50]. We classified the chemicals into three categories: 1) chemicals that had a <20% activity decrease in the single, high concentration screening, or had been assigned a 'hit-call' of 0 in the concentration-response AUR-TPO screening were classified as inactive in this assay; 2) chemicals with a 'hit-call' of 1 in AUR-TPO and a selectivity score greater than 1 were classified as active for TPO inhibition; and 3) chemicals with a 'hit-call' of 1 in AUR-TPO but with a selectivity score of 1 or less were classified as inconclusive for TPO inhibition.

The first dataset provided to the QSAR model developers at the National Food Institute (Food), Technical University of Denmark (DTU), consisted of structure information and experimental results for 1,126 ToxCast Phase I and II chemicals [34,40,41], including replicates, and was used for preparing a training set referred to as training set 1 (Figure 1). The second E1K dataset of an additional 771 chemicals from ToxCast [41,42], initially containing only structural information, was used for preparing a test set for external validation of the selected QSAR model build from training set 1 (see 2.3) (Figure 1). After determining the external validation statistics, the experimental results of the test set structures were made available to the model developers at DTU Food. The test set and training set 1 were then merged to form a second, larger training set referred to as training set 2 (Figure 1).

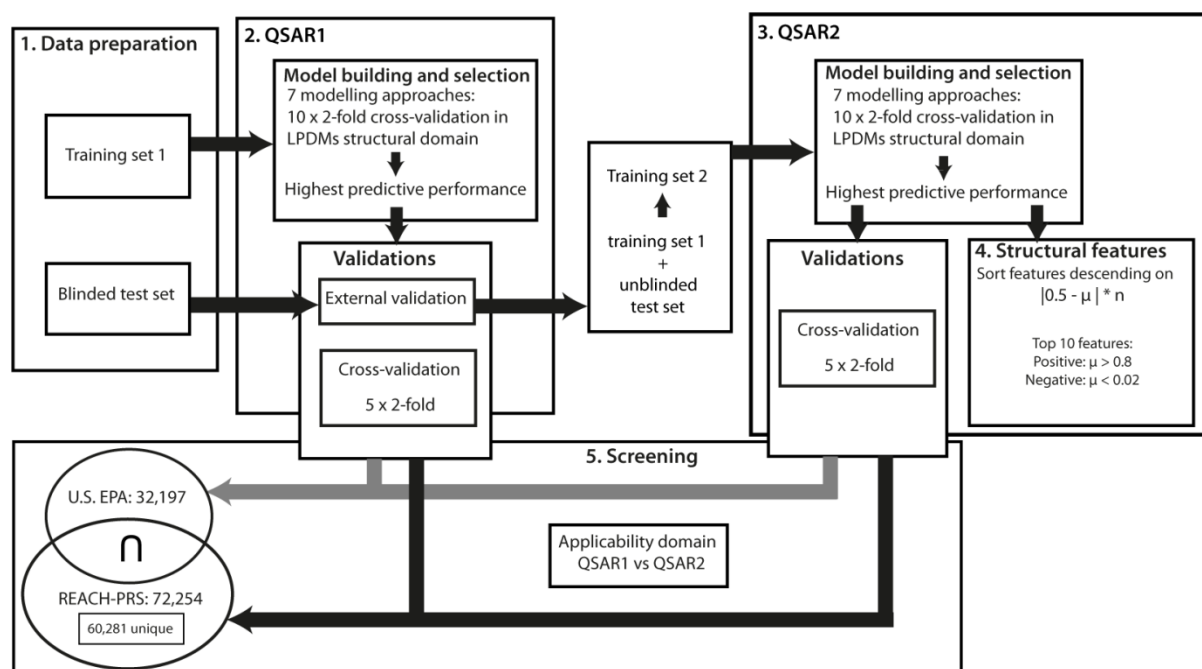


Figure 1. An overview of the datasets, modeling, structural feature sorting and screening. Here μ equals \bar{x} in the text and is the mean TPO inhibition experimental activity and n is the number of training set structures.

2.2 Structure Preparation

All chemical structures in the two U.S. EPA NCCT provided datasets had previously undergone an extensive quality control and structure curation procedure as part of the ToxCast program [41,51]. The QSAR software applied in this study handles organic chemical structures with an unambiguous 2D structure. We apply an overall definition of structures acceptable for QSAR processing in all our in-house QSAR software [45,46], as structures:

- containing at least two C atoms
- containing only the atoms H, Li, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and/or I; and,

- that are not mixtures consisting of two or more organic components

The structures that did not fulfill these criteria were removed from the two datasets. Further processing of the structural information included stripping off ions and neutralization of the organic parent structures, i.e. all structures were used in their non-ionized form.

Next, identical QSAR-ready structures within the first dataset were identified and their assigned experimental results were compared. For identical structures with concordant activities, only one of the structures was kept. If a group of identical structures had discrepant activities then the whole group was removed from the dataset. Next, structures with inconclusive experimental results, i.e. 'hit-call' of 1 in AUR-TPO and a selectivity score of 1 or less, were removed and the dataset now constituted training set 1 (Figure 1). The same duplicates removal procedure was performed by U.S. EPA NCCT scientists on the DTU Food experimentally-blinded E1K set, which then constituted the test set (Figure 1). Some of the QSAR-ready structures in the test set were identical to structures in training set 1 and were therefore excluded from the external validation. When the test set experimental results were made available to DTU Food, and training set 2 was prepared by merging the test set and training set 1 (Figure 1), the experimental results of the identified structural duplicates were compared. Again, if they had concordant experimental result only one of the structures was kept, while all the structures were removed in case of disagreement between the experimental results.

2.3 QSAR Modeling and Selection

We used the commercial software Leadscape® Predictive Data Miner (LPDM), a component of Leadscape® Enterprise Server version 3.2.4 [52], to build the QSAR models. Briefly, for each chemical structure in a training set LPDM automatically performs a systematic sub-structural analysis using a template library of more than 27,000 pre-defined structural features and calculates nine molecular descriptors (AlogP, Hydrogen Bond Acceptors and Donors, Lipinski Score, Molecular Weight, Parent Atom Number, Parent Molecular Weight, Polar Surface Area, Rotatable Bonds) [53]. The structural features and molecular descriptors are included in a default descriptor set. In addition, the user may call a functionality in LPDM to generate and add new training set-dependent structural features (scaffolds) to the descriptor set. The pre-defined structural features, added scaffolds and numeric molecular descriptors are included in an initial descriptor set. From the initial descriptor set, an automatic descriptor selection procedure in LPDM selects the top 30% descriptors according to Yates X^2 -test for a binary response variable. For the current training set 1 and 2 with binary response variables, predictive models were built using partial logistic regression (PLR) with further selection of descriptors in an iterative procedure, and selection of the optimum number of PLR

factors based on least predictive residual sum of squares. LPDM has the option of building composite models, a type of ensemble models, for training sets with an imbalanced distribution of actives and inactives. With this option a number of sub-models are created by specifying the desired ratio of actives to inactives per sub-model training set, so that each of the sub-models contains the smaller class and a sample of the bigger class. The positive prediction probability (see 2.4) for a query chemical from a composite model is defined as the average of the positive prediction probabilities of all sub-models having the test chemical in the applicability domain (AD) [54].

Multiple modeling approaches were applied in LPDM to build seven predictive models for TPO inhibition first using training set 1 (Figure 1):

- 1) single (i.e., non-composite)
- 2) single with scaffolds
- 3) single with scaffolds and a reduced set of structural features
- 4) composite
- 5) composite with scaffolds
- 6) composite with scaffolds and a reduced set of structural features
- 7) composite model combining model 3 and the sub-models from model 6

In 1 and 4, the descriptors were selected among the default descriptors, i.e. the molecular descriptors and the predefined structural features, and used to build a single model and a composite model, respectively. Next, scaffolds were generated in LPDM for the training set structures and added to the initial descriptor set, which subsequently was used for descriptor selection for models 2 and 5. In models 3 and 6, the scaffold-enriched descriptor set was reduced using a built-in function in LPDM (i.e., 'Remove most features – (removes less similar features)') that removed certain similar structural features before the descriptor selection. This step was employed to achieve a higher-quality set of fewer structural features, eliminate highly similar or redundant ones, and reduce the risk of overfitting. In model 7, the single model 3 and the sub-models from composite model 6 were combined to constitute a new composite model with equal weight of all its sub-models.

During model building all seven models underwent a ten times two-fold cross-validation by the LPDM algorithm. The algorithm transfers knowledge of the selected descriptor set from the parent model when building the cross-validation models, and we therefore do not use it for our measures of absolute predictive performance, but only to guide relative performance-based selection between the seven preliminary models. Among the seven predictive models built from training set 1, we selected the model with the highest performance from the LPDM cross-validation for further validation and screening studies (Figure 1). The selected model, called QSAR1, was then closed for further development (Figure 1).

2.4 Applicability Domain Definition

The definition of the AD applied in this project consists of two components: 1) the definition of a structural domain in LPDM, and 2) a DTU Food in-house class probability refinement on the output from LPDM:

- 1) For a test compound to be within LPDM's structural domain it was required that: all molecular descriptors used in the model could be calculated, it contained at least one structural feature used in the model, and it had at least 30% Tanimoto similarity with a training set compound [54]. The 30% Tanimoto similarity was a default cut-off in the LPDM software. For a test compound outside this structural domain no prediction call (active/inactive) was generated by LPDM. For test compounds within the LPDM structural domain, a positive prediction probability, p , between 0 and 1, was given together with the prediction call; actives having a $p \geq 0.5$ and inactives having a $p < 0.5$ [54].
- 2) To exclude less reliable predictions, i.e. those with a positive prediction probability close to the cutoff $p = 0.5$, we required $p \geq 0.7$ for active prediction calls and $p \leq 0.3$ for inactive prediction calls. Predictions within the LPDM structural domain but with an associated positive prediction probability in the interval 0.3 to 0.7 were thus defined as outside of the AD and excluded from the statistical analyses.

2.5 Validation of the Models

Next, the closed QSAR1 model underwent an external validation blinded to DTU Food using the test set to evaluate its predictive performance (Figure 1). U.S. EPA NCCT compared the DTU Food generated test set prediction calls within the AD (see 2.4) with the corresponding experimental results and calculated sensitivity, specificity, balanced accuracy and coverage. Sensitivity is the percentage of experimental actives correctly predicted, specificity is the percentage of the experimental inactives correctly predicted, and balanced accuracy is the average of the sensitivity and specificity [55]. The coverage is the proportion of test set compounds that had predictions within the model's AD.

The assigned experimental activities for the test set were then made available to DTU Food, who merged the test set with training set 1 to constitute the larger training set 2 (see 2.2). Training set 2 was used to build seven predictive models using the same modeling and LPDM cross-validation approaches described for training set 1 in 2.3, and of these the best performing model was selected (Figure 1). The selected model, called QSAR2, was closed for further development.

As described above, the LPDM cross-validation algorithm was, due to the issue with transfer of knowledge to the cross-validation models, only used to guide the selection of the best performing

model among the seven models built from training set 1 and 2, respectively. The two selected and closed models, QSAR1 and QSAR2, were each subsequently subjected to a DTU Food in-house five times two-fold stratified cross-validation procedure to further estimate their robustness and predictive performance (Figure 1). This was done by randomly removing 50% of the structures from the training set, preserving the ratio of actives and inactives. Then a cross-validation model was built on the reduced training set using the same modeling approach as the full, parent model, but without transferring any established information such as selected descriptors from the parent model. The cross-validation model was applied to predict the 50% of the training set that had been removed. Likewise, a cross-validation model was made using the removed 50% of the training set, and this model was used to predict the remaining 50%. This procedure was performed five times resulting in ten cross-validation models. Sensitivity, specificity and balanced accuracy were calculated for the in-AD predictions for each of the ten cross-validation models, and the mean and standard deviation (SD) were computed to give overall statistical measures of the predictive performance and robustness of the parent model based on the full-training set. The coverage, i.e. the mean percentage of how many of the predicted substances that had predictions within the AD of the ten cross-validation models, was also calculated.

2.6 Structural Features in QSAR2

To identify structural features in QSAR2 related to TPO inhibition or non-inhibition, respectively, all features in the model were sorted in descending order by:

$$|0.5 - \bar{x}| \cdot n$$

where n is the number of training set 2 structures containing the given feature, and \bar{x} is the mean TPO inhibition experimental activity (1 for actives and 0 for inactives) of the n training set structures.

With this metric the QSAR2 structural features that discriminate well between the two classes, i.e. actives and inactives, and are contained in the largest number of training set 2 structures are given the highest ranking. Based on this sorting, the top ten structural features with an $\bar{x} \geq 0.8$, i.e. structural features associated with activity, and an $\bar{x} \leq 0.02$, i.e. structural features associated with inactivity, respectively, were identified (Figure 1). The cutoff of $\bar{x} \leq 0.02$ was chosen instead of 0.2, which would have been symmetric to the $\bar{x} \geq 0.8$ cutoff for activity associated structural features, due to the larger proportion of inactive structures in the training set.

2.7 Screening Large Chemical Inventories

The structures in the REACH-PRS inventory were originally curated from deliverable 3.4 of the OpenTox EU project and had previously been processed through the structure preparation steps

described in 2.2 [56]. The 72,524 QSAR-ready REACH-PRS structures included structural duplicates, and the REACH-PRS set thus contained a total of 60,281 unique structures (Figure 1). The U.S. EPA inventory was also previously processed through the structure preparation steps described in 2.2 and 32,197 unique QSAR-ready structures remained. Both the REACH-PRS set and the U.S. EPA set were screened through the QSAR1 and QSAR2 TPO inhibition models to identify substances with the potential to inhibit TPO. We applied both QSAR1 and QSAR2 to be able to assess the effect of adding the test set structures to training set 2 with regard to the coverages of the two inventories and the prevalences of predicted TPO inhibitors. While QSAR2 is likely to provide better coverages of the inventories, the lack of an external validation of QSAR2 may for some purposes suggest that QSAR1 is a more appropriate model. The overlaps in substances as well as unique structures between U.S. EPA and REACH-PRS were identified (Figure 1). The proportion of the QSAR-predicted U.S. EPA and REACH-PRS substances within the AD of QSAR1 and QSAR2 and the activity distributions of the predictions were calculated.

3. Results and Discussion

This is to our knowledge the first study to develop global binary QSAR models for TPO inhibition and apply them to predict two large and structurally diverse chemical inventories containing man-made substances for their TPO inhibiting potential.

3.1 The Training and Test Sets

The number of QSAR-ready structures and the distribution of active and inactive experimental results in training set 1, the test set and training set 2 are summarized in Table 1 (will be made available in a supplementary file for submission). The numbers given in Table 1 reflect the situation after removing structures that were either unsuited for QSAR processing in the applied software, structural duplicates or had inconclusive experimental results. In training set 1 this resulted in the removal of 72 structures due to structural QSAR criteria, i.e. structures unacceptable for QSAR processing, 21 due to structural duplicates (four of these due to conflicting experimental results), and 156 due to inconclusive experimental results; in total 249 out of the 1,126 initial structure entries. In the external validation test set, a total of 125 out of the 771 initial E1K structure entries were removed; 14 due to structural QSAR criteria, 23 due to overlap with training set 1 structures, 14 due to internal structural duplicates (two of these due to conflicting experimental results), and 74 due to inconclusive experimental results. When merging training set 1 and the test set, which at this point was un-blinded to DTU Food, the experimental results of the 23 structures removed from the test set due to overlap with training set 1 structures were compared with their corresponding

training set 1 experimental results. In four cases the experimental results disagreed, and these structures were therefore removed from the final training set 2 (Table 1).

Table 1. Number of structures in the QSAR-ready training sets 1 and 2, and test set with the distribution of active and inactive experimental results for TPO inhibition.

Datasets	Total number of unique structures	Active (%)	Inactive (%)
Training set 1	877	130 (14.8)	747 (85.2)
Test set*	646	100 (15.5)	546 (84.5)
Training set 2**	1519	230 (15.1)	1289 (84.9)

*The experimental results of the test set were masked to DTU Food model developers until after being predicted in QSAR1. ** some of the training set 1 structures were tested again together with the test set structures, and of these four structures had different activities compared to the training set 1 activity. The four training set 1 structures were removed from training set 2.

The chemical structures in the provided datasets had undergone thorough quality control and curation [41,51]. In addition, since the datasets originated from the same source, i.e. U.S. EPA NCCT, and all chemicals had been screened in the same testing protocols and undergone the same data processing, this has likely contributed to a decrease in the experimental variability. The data in training set 1 and 2 and the test set were therefore assessed to be of high quality [34,39] and expected to be a good basis for QSAR model development. The quality of the AUR-TPO assay has been assessed previously [34,39], which indicated excellent performance and intralaboratory repeatability (r_z' from 0.77 to 0.83 and rCV of 3–4%). The AUR-TPO assay measures the fluorescence intensity from the commercial peroxidase substrate, Amplex®UltraRed (AUR), which is converted to Amplex UltroxRed by a peroxidase in the presence of hydrogen peroxide. A decrease in fluorescence intensity in response to a chemical is an indirect measure of TPO inhibition. The reaction chemistry and oxidation product of AUR is proprietary and the exact reaction(s) inhibited and its reversibility cannot be identified [34]. Therefore, the AUR-TPO assay read out has multiple potential confounders, including: non-specific enzyme inhibition; reactive, autofluorescent or fluorescence quenching chemicals; and other sources of interference with the peroxidase reaction [34,39]. When comparing results from the AUR-TPO assay with results from the lower throughput orthogonal guaiacol oxidation assay, the AUR-TPO assay was previously found to have a sensitivity of 86% and a specificity of 39% [34]. Part of the high sensitivity of AUR-TPO could be due to a higher rate of false positive results from confounding non-specific activity decrease, a known problem with loss-of signal assays. Identification and removal of such potentially AUR-TPO false positive TPO inhibitors in the datasets was attempted by the application of the selectivity score filter [34] and the inconclusive category, i.e. AUR-TPO positives with a selectivity score less than 1, see section 2.1. However, not all mechanisms potentially causing non-specific activity decrease, e.g. fluorescence quenching, have been addressed in the selectivity score [34] and so the presence of false positive TPO inhibitors in the training and test sets cannot be excluded. Furthermore, the tiered screening approach in AUR-TPO with a cutoff of 20% activity decrease in the initial single, high-concentration screening [34] may

have produced some false negatives as it cannot be excluded that a portion of the chemicals causing an activity decrease below the cutoff would have been positive if screened for concentration-response. In addition to the potential confounding effects in the raw experimental outputs, the models applied for the 'hit-call' assignment and the selectivity score algorithm are also subject to some degree of uncertainty in their results.

3.2 QSAR Modeling and Selection

Table 2 shows the LPDM cross-validation results for the seven models built from training set 1 and 2, respectively. As mentioned above, the LPDM cross-validation was used to guide relative performance-based selection between the seven preliminary models. As can be seen in Table 2, the composite models 4 to 7 outperformed the single models 1, 2 and 3 in the LPDM cross-validation with regard to the balanced accuracy (Table 2). This is most likely an effect of the imbalanced distribution of actives and inactives in both training sets with a ratio of approximately 1:6 (Table 1). The composite model option in LPDM was implemented to handle such imbalanced training sets to include also a high proportion of the bigger class and thereby optimize the size of the AD [54].

Table 2. The results from the LPDM cross-validation of the seven built models from training set 1 and 2, respectively.

LPDMs 10 times two-fold cross-validation results							
Model							
Training set 1	Sensitivity (%)	Specificity (%)	Balanced accuracy (%)	TP*	FP*	TN*	FN*
1	43.0	96.8	69.9	49	21	626	65
2	48.2	96.0	72.1	55	26	621	59
3	50.0	96.3	73.2	57	24	623	57
4	72.9	82.7	77.8	94	105	502	35
5	81.4	78.2	79.8	105	136	498	24
6	84.5	80.3	82.4	109	123	502	20
7	74.6	92.5	83.6	97	55	676	33
Training set 2	Sensitivity (%)	Specificity (%)	Balanced accuracy (%)	TP*	FP*	TN*	FN*
1	46.5	96.9	71.2	99	40	1153	114
2	49.8	96.1	73.0	106	46	1147	107
3	46.5	96.7	71.6	99	39	1154	114
4	79.1	79.9	79.5	182	233	928	48
5	75.7	79.5	77.6	174	240	931	56
6	76.1	78.4	77.3	175	253	918	55
7	71.3	92.6	82.0	164	95	1187	66

*TP: true positives, FP: false positives, TN: true negatives, FN: false negatives. The numbers are averages of the ten iterations as given by LPDM.

In this work we employed a new approach where a single, unbalanced model (i.e., model 3) was added as a sub-model, together with the balanced sub-models from a composite model (i.e., model 6), to form a new composite model (i.e., model 7). This addition caused a significant reduction in the number of false positive (FP) predictions produced in the LPDM cross-validation as compared to

model 6 alone (see Table 2). For both training set 1 and 2 this resulted in a remarkable increase in the LPDM cross-validation specificity while causing a smaller reduction in sensitivity (Table 2), and together this explains why model 7, in both cases, outperformed the other composite models 4, 5 and 6. To conclude, model 7 was the best performing among the seven models for both training set 1 and 2, and therefore selected for both training sets, and these models were named QSAR1 and QSAR2, respectively (Table 3).

Table 3. Modeling approach applied and the predictive performances for QSAR1 and QSAR2.

Model	Statistical Parameter	Cross-Validation*, % (SD, %)	External Validation**, % (actual numbers)
QSAR1 Approach 7 Sub-models: 7	Sensitivity	72.3 (10.1)	79.7 (47/(47 + 12))
	Specificity	89.0 (2.8)	90.8 (266/(266 + 27))
	Balanced accuracy	80.6 (4.6)	85.3 ((79.7 + 90.8)/2)
	Coverage	51.6 (4.7)	54.5 (352/646)
QSAR2 Approach 7 Sub-models: 7	Sensitivity	75.6 (5.0)	-
	Specificity	89.8 (1.5)	-
	Balanced accuracy	82.7 (2.2)	-
	Coverage	57.8 (5.4)	-

*A five times two-fold cross-validation, ** A blinded external validation with the experimental results of the test set being masked to the model developers at DTU Food.

3.3 Predictive Performance of the QSAR Models

The two selected and final models, QSAR1 and QSAR2, underwent a five times two-fold DTU Food in-house cross-validation procedure to evaluate their predictive performance and robustness. QSAR1 also underwent a DTU Food blinded external validation with the test set. The results from the validation studies are presented in Table 3 and demonstrate high predictive performance, i.e. balanced accuracies of 85.3% by external validation for QSAR1 and 82.7% by cross-validation for QSAR2, respectively.

Adding the test set to training set 1 to build QSAR2 served multiple purposes. One purpose was to explore how much the added test set would enlarge the AD of the model and thereby increase the coverages of the two large chemical screening inventories, U.S. EPA and REACH-PRS. The coverage of QSAR2 was roughly 6% larger in the cross-validation (Table 3) and 10% larger for both screening inventories (Table 5) than the respective coverages of QSAR1. A second purpose of adding the test set in QSAR2 was to explore the possible improvements in predictive performance. To do this, we first built the smaller QSAR1 model and performed both a rigorous five times two-fold cross-validation procedure and a large external validation with the test set. As can be seen in Table 3 the validation procedures show that QSAR1 has high predictive performance and is a robust model, i.e. a balanced accuracy of 85.3% in external validation and 80.6% with an SD of 4.6% in the cross-validation. A comparison of the statistical parameters from the two validation methods indicates that the rigorous cross-validation procedure applied does not overestimate the model's predictive

performance, but rather, outputs conservative estimates. This conservative nature of the cross-validation is likely due to the rigorous procedure of removing 50% of the full training set to build the cross-validation models. Such a procedure is especially hard on the proportionally few actives in training set 1, i.e. 130 out of 877 (Table 1), which is also reflected in the relatively high SD of 10% in the sensitivity of the ten QSAR1 cross-validation models as well as its lower mean value (72.3%) compared to the sensitivity from the external validation (79.7%) (Table 3). The structures in the test set used for the DTU-blinded external validation of QSAR1 were not selected due to specific TPO inhibition concerns or to serve as a representative test set for QSAR1, but instead selected because they are included in the U.S. EPA regulatory ToxCast universe based on potential for exposure, and not because of prior concern about endocrine disruptive effects [41,42].

The procedure of performing both independent and robust cross-validation and a large, representative and prospective external validation is optimal when evaluating a model's predictive performance, but external validation has the disadvantage of withholding what may be valuable data from the model itself. Adding all available data to a training set can, in addition to expanding the AD, also result in a model with a higher predictive performance, depending on the characteristics of the added data. The QSAR2 model could not undergo an external validation procedure due to lack of another external test set. Previous studies have shown that robust cross-validations give reliable estimates of a model's predictive performance (e.g. [57,58]). This, together with the results from the cross-validation vs. external validation results of QSAR1, suggests that the applied cross-validation procedure can be used for assessing QSAR2's predictive performance. Due to the conservative nature of the two-fold cross-validation, we anticipate that QSAR2 will have a similar or higher predictive performance if it underwent a large external validation with a test set generated using the same protocol and data processing. As can be seen from Table 3, the cross-validation sensitivity was slightly increased in QSAR2 (75.6%) compared to QSAR1 (72.3%) and the sensitivity SD was reduced from 10.1% to 5%. This is most likely the effect of an increase in actives from 130 in training set 1 to 230 in training set 2, which renders the 50% exclusion in the cross-validation procedure less influential on the sensitivity. As there were already many inactives in training set 1, the addition of more inactives to training set 2 did, as expected, not have the same high impact on the specificity, which went from 89.0% (SD = 2.8%) in QSAR 1 to 89.8% (SD = 1.5%) in QSAR2.

3.4 Top Structural Features in QSAR2

The ten most frequent and discriminating predictive structural features associated with actives and inactives, respectively, in QSAR2 are shown in Figure 2. Among the highest ranking structural features associated with activity were versions of phenols, anisole and aniline. The most frequent

structural features associated with inactivity included ethers, esters, aryl halides and a tertiary amine. To our knowledge structural docking or pharmacophore studies for TPO have not been performed (Simmons *et al.*, in prep).

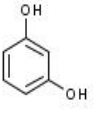

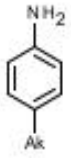
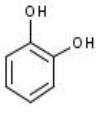
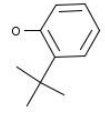
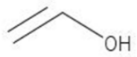
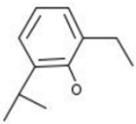
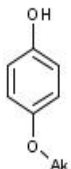
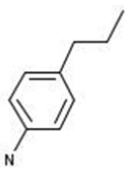
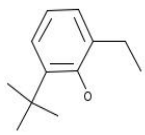
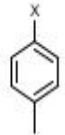
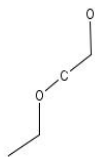
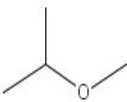
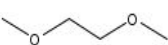
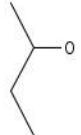
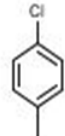
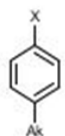
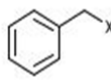
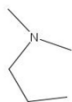
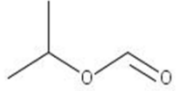
13/0  benzene, 1,3-dihydroxy-	13/2  Scaffold 288	11/1  benzene, 1-alkyl-,4-amino(NH2)-	9/0  benzene, 1,2-dihydroxy-	9/2  Scaffold 297
6/0  alcohol, alkenyl-	7/1  Scaffold 576	5/0  benzene, 1-alkoxy-,4-hydroxy-	5/0  Scaffold 306	6/1  Scaffold 574
0/71  Scaffold 110	1/62  Scaffold 342	1/57  Scaffold 210	0/52  Scaffold 253	0/49  Scaffold 303
0/47  Scaffold 108	0/44  benzene, 1-alkyl-,4-halo-	0/41  halide, benzyl-	0/36  Scaffold 454	0/35  Scaffold 194

Figure 2. The structural features used in QSAR2 were sorted on $|0.5 - \bar{x}(\text{TPO inhibition activity})| \cdot n$, and the ten most frequent and discriminating structural features alerting for activity ($\bar{x}(\text{TPO inhibition activity}) \geq 0.8$) and inactivity ($\bar{x}(\text{TPO inhibition activity}) \leq 0.02$) are shown here. Ak matches saturated carbon and X matches the halogen atoms Cl, Br, I or F. Numbers in the upper left corners display the ratio of TPO inhibitors/non-inhibitors in training set 2 for the specific structural feature.

3.5 The Screening Results

We found a total of 27,444 substances present in both the U.S. EPA and the full REACH-PRS inventories. There were 19,279 unique structures in common in the two inventories (Table 4). To our knowledge this is the first study that has quantified the overlap between these two inventories, both with regard to overall substances and unique structures. The high overlap between the U.S. EPA set

and the REACH-PRS set was not surprising since both inventories represent collections of man-made, environmental chemicals in the U.S. and EU, respectively.

Table 4. The overlap in substances and unique structures between the U.S. EPA and REACH-PRS inventories.

Overlap analysis	U.S. EPA*	REACH-PRS**	Total number	In common		Unique to a set	
				REACH-PRS in U.S. EPA	U.S. EPA in REACH-PRS	REACH-PRS	U.S. EPA
Structure entries	32,197	72,524	104,721	27,444	19,279	45,080	12,918
Unique structures	32,197	60,281	92,478	19,279	19,279	41,002	12,918

*U.S. EPA: QSAR-ready structures from an U.S. EPA selected inventory of man-made chemical structures to which humans are potentially exposed, ** REACH-PRS: QSAR-ready structures from the REACH pre-registered substances list

Both the U.S. EPA and REACH-PRS inventories were screened using QSAR1 and QSAR2 for TPO inhibition. In Table 5 the coverage of the two substance inventories, i.e. the proportion of the full set predicted within the AD of the model, and the number of active and inactive predictions are presented for each model. As mentioned earlier, the coverages of QSAR2 was as expected larger than QSAR1 of both screening sets. The percentage of chemicals in the two inventories with active predictions in the AD of the two models ranged from 16.5% to 19.3% (Table 5), which was slightly higher than the percentage of experimentally determined actives of 14.8% to 15.5% in the training and test sets (Table 1).

Table 5. The coverage (AD) and the number of active/inactive predictions of the U.S. EPA and REACH-PRS inventories in QSAR1 and QSAR2.

	Total	QSAR 1			QSAR2		
		In AD (%)	Active (%)	Inactive (%)	In AD (%)	Active (%)	Inactive (%)
U.S. EPA*	32,197	16,898 (52.5)	2855 (16.9)	14,043 (83.1)	19,392 (60.2)	3201 (16.5)	16,191 (83.5)
REACH-PRS**	72,524	38,661 (53.3)	7,128 (18.4)	31,533 (81.6)	45,540 (62.8)	8,790 (19.3)	36,750 (80.7)
REACH-PRS unique	60,281	32,334 (53.6)	5,879 (18.2)	26,455 (81.8)	37,784 (62.7)	7,166 (19.0)	30,618 (81.0)

*U.S. EPA: QSAR-ready structures from an U.S. EPA selected inventory of man-made chemical structures to which humans are potentially exposed, ** REACH-PRS: QSAR-ready structures from the REACH pre-registered substances list

As mentioned earlier, the chemicals in the experimental datasets were not selected on the basis of expected TPO inhibition effects. It is not known to what extent these slightly higher percentages of TPO inhibitors in the two predicted screening sets are due to FP predictions or if they reflect a true TPO inhibitor prevalence. The validation studies showed that both QSAR1 and QSAR2 have specificities >10% higher than their respective sensitivities (Table 3), and therefore both models are expected to, in a balanced universe, make relatively more FN than FP predictions.

3.6 Butylated Hydroxyanisole as a Potential Thyroid Hormone Disruptor

We searched the two chemical inventories for possible examples of human-relevant chemicals with known indications for adverse neurodevelopmental outcomes. Included in both the U.S. EPA and the REACH-PRS set were the two isomers of butylated hydroxyanisole (BHA, CASN 25013-16-5), 2-*tert*-Butyl-4-hydroxyanisole (2-BHA, CASN 88-32-4) and 3-*tert*-Butyl-4-hydroxyanisole (3-BHA, CASN 121-00-6) (Figure 3).

BHA is manufactured and/or imported to the EU in a total of 100-1,000 tonnes per year and is used as an antioxidant and preservative in e.g. food, food contact materials, cosmetics, and pharmaceuticals [59–61]. It is an anticipated human carcinogen [62] and is has been noted to have published evidence of developmental neurotoxicity (DNT) in mammals [63,64]. Both *in vitro* and *in vivo* published studies indicate that the BHA isomers have endocrine-modulating potential, with most evidence for estrogenic and androgenic effects [61,65–70]. Based on this, BHA is on both the EU list of potential endocrine disruptors [71,72] and on the SIN (Substitute It Now!) List [73,74]. However, more data is needed to fully elucidate BHA's potential as an endocrine disruptor and its mode of action(s) in DNT [61].

2-*tert*-Butyl-4-hydroxyanisole (2-BHA)

3-*tert*-Butyl-4-hydroxyanisole (3-BHA)*



9/2 Scaffold 297	5/0 benzene, 1-alkoxy-,4-hydroxy-	3/0 benzene, 1-hydroxy-,4-methoxy-
-------------------------	--	---

Figure 3. The two isomers of BHA and the three predictive structural features alerting for activity in QSAR2 selected based on highest $|0.5 - \bar{x}(\text{TPO inhibition activity})| * n$ and an $\bar{x} \geq 0.8$. *3-BHA (CASN 121-00-6) was included in the training set and is the closest analog to 2-BHA (CASN 88-32-4).

Both 2- and 3-BHA were predicted active for TPO inhibition by QSAR2, and 3-BHA was included in the QSAR2 training set as a TPO inhibitor. Studies in rats and pigs indicate that exposure to BHA (mixture of the two isomers) *in utero* can cause effects such as changed T4 serum levels, altered thyroid gland function and histology, and altered brain weight and behavior in the offspring

[64,65,70]. TPO inhibition is as mentioned above identified to be the MIE in an AOP for thyroid-related neurodevelopmental adverse effects (under development) [41]. The three common top activity-associated structural features from QSAR2 in the two isomers were identified as described in 2.6 and are shown in Figure 3. Two of the features, “Scaffold 297” and “benzene, 1-alcoxy-,4-hydroxy” were among the top ten structural features associated with activity in QSAR2 (Figure 2). “Scaffold 297” was present in eleven training set 2 structures of which nine were experimentally active for TPO inhibition. The “benzene, 1-alcoxy-,4-hydroxy” structural feature was present in five training set 2 structures that were all experimentally active.

The QSAR2 training set including flags for the test set structures of QSAR1 will be made available in the supplementary material. Work is underway to make the training sets available from the U.S. EPA ToxCast website. Furthermore, predictions for around 640,000 structures in QSAR2, including the 72,524 REACH-PRS structures, will be made available from the online Danish (Q)SAR Database [46]. QSAR2 will also be made available for prediction of user-submitted structures in a coming free online Danish (Q)SAR Models sister-site to the Danish (Q)SAR database at the DTU homepage [46].

4. Conclusions

The present study reports the development, validation, and application of two global, binary composite QSAR models for TPO inhibition *in vitro*. The first model, QSAR1, showed high predictive performance in both cross-and external validation with balanced accuracies of 80.6% (SD = 4.6%) and 85.3%, respectively. QSAR2, the second model enlarged with the external test set of QSAR1, showed improved robustness and predictive performance in cross-validation compared to QSAR1, i.e. a balanced accuracy of 82.7% (SD = 2.2%), and this was largely driven by an increase in sensitivity from 72.3% (SD = 10.1%) of QSAR1 to 75.6% (SD = 5.0%) of QSAR2. The top-ten structural features in QSAR2 related to TPO inhibition and non-inhibition, respectively, were identified. The two QSAR models were used to screen two large chemical inventories from the U.S. and EU containing structurally diverse man-made chemicals to which humans are potentially exposed. QSAR2 showed an increase in coverage of around 10% for both inventories relative to QSAR1, and of the substances predicted within QSAR2’s AD, 8,790 (19.3%) REACH-PRS substances and 7,166 (19.0%) U.S. EPA substances, respectively, were predicted to be TPO inhibitors. Among the predicted TPO inhibitors were the two isomers of BHA, which have previously been shown to cause both TH and neurological effects in animal studies. These QSAR predictions may contribute to elucidating the mode of action by which BHA results in these altered TH levels and neurological outcomes. Overall, predictions from the two models can be used to prioritize chemicals for further testing in considerations of possible

concerns for downstream adverse outcomes (e.g., DNT) [75,76]. They may also be used e.g. in read-across cases or in IATA WoE assessments.

Conflict of Interest Statement

The authors declare that they have no conflict of interest in relation with this paper.

Acknowledgements

We would like to thank the Danish 3R Center and the Danish Environmental Protection Agency for supporting the project.

References

- [1] G.R. Williams, Neurodevelopmental and Neurophysiological Actions of Thyroid Hormone, *J. Neuroendocrinol.* 20 (2008) 784–794. doi:10.1111/j.1365-2826.2008.01733.x.
- [2] P.M. Yen, Physiological and molecular basis of thyroid hormone action., *Physiol. Rev.* 81 (2001) 1097–1142. <http://www.ncbi.nlm.nih.gov/pubmed/11427693>.
- [3] R.T. Zoeller, S.W. Tan, R.W. Tyl, General Background on the Hypothalamic-Pituitary-Thyroid (HPT) Axis, *Crit. Rev. Toxicol.* 37 (2007) 11–53. doi:10.1080/10408440601123446.
- [4] R.T. Zoeller, K.M. Crofton, Mode of Action: Developmental Thyroid Hormone Insufficiency—Neurological Abnormalities Resulting From Exposure to Propylthiouracil, *Crit. Rev. Toxicol.* 35 (2005) 771–781. doi:10.1080/10408440591007313.
- [5] E. Cuevas, E. Ausó, M. Telefont, G.M. de Escobar, C. Sotelo, P. Berbel, Transient maternal hypothyroxinemia at onset of corticogenesis alters tangential migration of medial ganglionic eminence-derived neurons, *Eur. J. Neurosci.* 22 (2005) 541–551. doi:10.1111/j.1460-9568.2005.04243.x.
- [6] K.L. Howdeshell, A Model of the Development of the Brain as a Construct of the Thyroid System, *Environ. Health Perspect.* 110 (2002) 337–348. doi:10.1289/ehp.02110s3337.
- [7] J. Kratzsch, F. Pulzer, Thyroid gland development and defects, *Best Pract. Res. Clin. Endocrinol. Metab.* 22 (2008) 57–75. doi:10.1016/j.beem.2007.08.006.
- [8] M.D. Miller, K.M. Crofton, D.C. Rice, R.T. Zoeller, Thyroid-Disrupting Chemicals: Interpreting Upstream Biomarkers of Adverse Outcomes, *Environ. Health Perspect.* 117 (2009) 1033–1041. doi:10.1289/ehp.0800247.
- [9] P. Berbel, J.L. Mestre, A. Santamaría, I. Palazón, A. Franco, M. Graells, A. González-Torga, G.M. de Escobar, Delayed Neurobehavioral Development in Children Born to Pregnant Women with Mild Hypothyroxinemia During the First Month of Gestation: The Importance of Early Iodine Supplementation, *Thyroid.* 19 (2009) 511–519. doi:10.1089/thy.2008.0341.
- [10] K.M. Crofton, Developmental Disruption of Thyroid Hormone: Correlations with Hearing Dysfunction in Rats, *Risk Anal.* 24 (2004) 1665–1671. doi:10.1111/j.0272-4332.2004.00557.x.
- [11] E.S. Goldey, L.S. Kehn, G.L. Rehnberg, K.M. Crofton, Effects of Developmental Hypothyroidism on Auditory and Motor Function in the Rat, *Toxicol. Appl. Pharmacol.* 135 (1995) 67–76. doi:10.1006/taap.1995.1209.
- [12] L. Kooistra, S. Crawford, A.L. van Baar, E.P. Brouwers, V.J. Pop, Neonatal Effects of Maternal Hypothyroxinemia During Early Pregnancy, *Pediatrics.* 117 (2006) 161–167. doi:10.1542/peds.2005-0227.

- [13] Y. Li, Z. Shan, W. Teng, X. Yu, Y. Li, C. Fan, X. Teng, R. Guo, H. Wang, J. Li, Y. Chen, W. Wang, M. Chawinga, L. Zhang, L. Yang, Y. Zhao, T. Hua, Abnormalities of maternal thyroid function during pregnancy affect neuropsychological development of their children at 25-30 months, *Clin. Endocrinol. (Oxf)*. 72 (2010) 825–829. doi:10.1111/j.1365-2265.2009.03743.x.
- [14] G. Morreale de Escobar, M. Jesús Obregón, F. Escobar del Rey, Is Neuropsychological Development Related to Maternal Hypothyroidism or to Maternal Hypothyroxinemia? 1, *J. Clin. Endocrinol. Metab.* 85 (2000) 3975–3987. doi:10.1210/jcem.85.11.6961.
- [15] V.J. Pop, E.P. Brouwers, H.L. Vader, T. Vulsma, A.L. van Baar, J.J. de Vijlder, Maternal hypothyroxinaemia during early pregnancy and subsequent child development: a 3-year follow-up study, *Clin. Endocrinol.* 59 (2003) 282–288. doi:10.1046/j.1365-2265.2003.01822.x.
- [16] V.J. Pop, J.L. Kuijpers, A.L. van Baar, G. Verkerk, M.M. van Son, J.J. de Vijlder, T. Vulsma, W.M. Wiersinga, H.A. Drexhage, H.L. Vader, Low maternal free thyroxine concentrations during early pregnancy are associated with impaired psychomotor development in infancy, *Clin. Endocrinol. (Oxf)*. 50 (1999) 149–155. doi:10.1046/j.1365-2265.1999.00639.x.
- [17] R.T. Zoeller, J. Rovet, Timing of Thyroid Hormone Action in the Developing Brain: Clinical Observations and Experimental Findings, *J. Neuroendocrinol.* 16 (2004) 809–818. doi:10.1111/j.1365-2826.2004.01243.x.
- [18] J.E. Haddow, G.E. Palomaki, W.C. Allan, J.R. Williams, G.J. Knight, J. Gagnon, C.E. O’Heir, M.L. Mitchell, R.J. Hermos, S.E. Waisbren, J.D. Faix, R.Z. Klein, Maternal Thyroid Deficiency during Pregnancy and Subsequent Neuropsychological Development of the Child, *N. Engl. J. Med.* 341 (1999) 549–555. doi:10.1056/NEJM199908193410801.
- [19] E.N. Pearce, Thyroid hormone and obesity, *Curr. Opin. Endocrinol. Diabetes Obes.* 19 (2012) 408–413. doi:10.1097/MED.0b013e328355cd6c.
- [20] C. Wang, The Relationship between Type 2 Diabetes Mellitus and Related Thyroid Diseases, *J. Diabetes Res.* 2013 (2013) 1–9. doi:10.1155/2013/390534.
- [21] K.L. Dionisio, A.M. Frame, M.-R. Goldsmith, J.F. Wambaugh, A. Liddell, T. Cathey, D. Smith, J. Vail, A.S. Ernstoff, P. Fantke, O. Jolliet, R.S. Judson, Exploring consumer exposure pathways and patterns of use for chemicals in the environment, *Toxicol. Reports.* 2 (2015) 228–237. doi:10.1016/j.toxrep.2014.12.009.
- [22] P.P. Egeghy, R. Judson, S. Gangwal, S. Mosher, D. Smith, J. Vail, E.A. Cohen Hubal, The exposure data landscape for manufactured chemicals, *Sci. Total Environ.* 414 (2012) 159–166. doi:10.1016/j.scitotenv.2011.10.046.
- [23] R. Judson, A. Richard, D.J. Dix, K. Houck, M. Martin, R. Kavlock, V. Dellarco, T. Henry, T. Holderman, P. Sayre, S. Tan, T. Carpenter, E. Smith, The Toxicity Data Landscape for Environmental Chemicals, *Environ. Health Perspect.* 117 (2009) 685–695. doi:10.1289/ehp.0800168.
- [24] M.-R. Goldsmith, C.M. Grulke, R.D. Brooks, T.R. Transue, Y.M. Tan, A. Frame, P.P. Egeghy, R. Edwards, D.T. Chang, R. Tornero-Velez, K. Isaacs, A. Wang, J. Johnson, K. Holm, M. Reich, J. Mitchell, D.A. Vallero, L. Phillips, M. Phillips, J.F. Wambaugh, R.S. Judson, T.J. Buckley, C.C. Dary, Development of a consumer product ingredient database for chemical exposure screening and prioritization, *Food Chem. Toxicol.* 65 (2014) 269–279. doi:10.1016/j.fct.2013.12.029.
- [25] A.J. Murk, E. Rijntjes, B.J. Blaauboer, R. Clewell, K.M. Crofton, M.M.L. Dingemans, J. David Furlow, R. Kavlock, J. Köhrle, R. Opitz, T. Traas, T.J. Visser, M. Xia, A.C. Gutleb, Mechanism-based testing strategy using in vitro approaches for identification of thyroid hormone disrupting chemicals, *Toxicol. Vitro.* 27 (2013) 1320–1346. doi:10.1016/j.tiv.2013.02.012.

- [26] K.M. Crofton, E.S. Craft, J.M. Hedge, C. Gennings, J.E. Simmons, R.A. Carchman, W.H. Carter Jr., M.J. DeVito, Thyroid-Hormone–Disrupting Chemicals: Evidence for Dose-Dependent Additivity or Synergism, *Environ. Health Perspect.* 113 (2005) 1549–1554. doi:10.1289/ehp.8195.
- [27] R.L. Divi, D.R. Doerge, Mechanism-Based Inactivation of Lactoperoxidase and Thyroid Peroxidase by Resorcinol Derivatives, *Biochemistry.* 33 (1994) 9668–9674. doi:10.1021/bi00198a036.
- [28] M. V. Kirthana, F. Nawaz Khan, P.M. Sivakumar, M. Doble, P. Manivel, K. Prabakaran, V. Krishnakumar, Antithyroid agents and QSAR studies: inhibition of lactoperoxidase-catalyzed iodination reaction by isochromene-1-thiones, *Med. Chem. Res.* 22 (2013) 4810–4817. doi:10.1007/s00044-013-0475-x.
- [29] OECD, Proposal for a template, and guidance on developing and assessing the completeness of adverse outcome pathways, (2012). <http://www.oecd.org/chemicalsafety/testing/49963554.pdf> (accessed January 13, 2017).
- [30] AOP-Wiki, The AOP-Wiki homepage, (2017). <https://aopwiki.org/> (accessed March 13, 2017).
- [31] N.C. Kleinstreuer, K. Sullivan, D. Allen, S. Edwards, D.L. Mendrick, M. Embry, J. Matheson, J.C. Rowlands, S. Munn, E. Maull, W. Casey, Adverse Outcome Pathways: From Research to Regulation Scientific Workshop Report, *Regul. Toxicol. Pharmacol.* 76 (2016) 39–50. doi:10.1016/j.yrtph.2016.01.007.
- [32] OECD, Workshop on Integrated Approaches to Testing and Assessment, 2008. [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2008\)10&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2008)10&doclanguage=en) (accessed January 13, 2017).
- [33] K.E. Tollefsen, S. Scholz, M.T. Cronin, S.W. Edwards, J. de Knecht, K. Crofton, N. Garcia-Reyero, T. Hartung, A. Worth, G. Patlewicz, Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA), *Regul. Toxicol. Pharmacol.* 70 (2014) 629–640. doi:10.1016/j.yrtph.2014.09.009.
- [34] K. Paul Friedman, E.D. Watt, M.W. Hornung, J.M. Hedge, R.S. Judson, K.M. Crofton, K.A. Houck, S.O. Simmons, Tiered High-Throughput Screening Approach to Identify Thyroperoxidase Inhibitors Within the ToxCast Phase I and II Chemical Libraries, *Toxicol. Sci.* 151 (2016) 160–180. doi:10.1093/toxsci/kfw034.
- [35] AOPs, AOPs in AOP-Wiki as of March 2017, (2017). <https://aopwiki.org/aops> (accessed March 13, 2017).
- [36] AOP-42, Inhibition of Thyroperoxidase and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/42> (accessed March 13, 2017).
- [37] R.S. Fortunato, E.C. Lima de Souza, R.A. Hassani, M. Boufraquech, U. Weyemi, M. Talbot, O. Lagente-Chevallier, D.P. de Carvalho, J.-M. Bidart, M. Schlumberger, C. Dupuy, Functional Consequences of Dual Oxidase-Thyroperoxidase Interaction at the Plasma Membrane, *J. Clin. Endocrinol. Metab.* 95 (2010) 5403–5411. doi:10.1210/jc.2010-1085.
- [38] J. Ruf, P. Carayon, Structural and functional aspects of thyroid peroxidase, *Arch. Biochem. Biophys.* 445 (2006) 269–277. doi:10.1016/j.abb.2005.06.023.
- [39] K.B. Paul, J.M. Hedge, D.M. Rotroff, M.W. Hornung, K.M. Crofton, S.O. Simmons, Development of a Thyroperoxidase Inhibition Assay for High-Throughput Screening, *Chem. Res. Toxicol.* 27 (2014) 387–399. doi:10.1021/tx400310w.
- [40] D.J. Dix, K.A. Houck, M.T. Martin, A.M. Richard, R.W. Setzer, R.J. Kavlock, The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals, *Toxicol. Sci.* 95 (2007) 5–

12. doi:10.1093/toxsci/kfl103.
- [41] A.M. Richard, R.S. Judson, K.A. Houck, C.M. Grulke, P. Volarath, I. Thillainadarajah, C. Yang, J. Rathman, M.T. Martin, J.F. Wambaugh, T.B. Knudsen, J. Kancherla, K. Mansouri, G. Patlewicz, A.J. Williams, S.B. Little, K.M. Crofton, R.S. Thomas, ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology, *Chem. Res. Toxicol.* 29 (2016) 1225–1251. doi:10.1021/acs.chemrestox.6b00135.
- [42] EDSP21 Work Plan, The Incorporation of In Silico Models and In Vitro High Throughput Assays in the Endocrine Disruptor Screening Program (EDSP) for Prioritization and Screening, (2011). https://www.epa.gov/sites/production/files/2015-07/documents/edsp21_work_plan_summary_overview_final.pdf (accessed March 13, 2017).
- [43] ECHA, Guidance on information requirements and chemical safety assessment - Chapter R.6: QSARs and grouping of chemicals, (2008). https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf (accessed March 16, 2017).
- [44] OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, 2 (2007) 1–154. [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2007\)2&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en) (accessed December 8, 2016).
- [45] QSAR, User Manual for the Danish (Q)SAR Database, (2015). http://qsar.db.food.dtu.dk/Danish_QSAR_Database_Draft_User_manual.pdf (accessed March 28, 2017).
- [46] QSARDB, Danish (Q)SAR Database, (2015). <http://qsar.food.dtu.dk/> (accessed March 14, 2017).
- [47] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A.M. Richard, C.M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E.B. Wedebye, F. Grisoni, G.F. Mangiatordi, G.M. Incisivo, H. Hong, H.W. Ng, I. V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N.G. Nikolov, O. Nicolotti, P.L. Andersson, Q. Zang, R. Politi, R.D. Beger, R. Todeschini, R. Huang, S. Farag, S.A. Rosenberg, S. Slavov, X. Hu, R.S. Judson, CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ. Health Perspect.* 124 (2016) 1023–1033. doi:10.1289/ehp.1510267.
- [48] A.M. Richard, C.R. Williams, Distributed structure-searchable toxicity (DSSTox) public database network: a proposal, *Mutat. Res. Mol. Mech. Mutagen.* 499 (2002) 27–52. doi:10.1016/S0027-5107(01)00289-5.
- [49] Z.A. Collier, K.A. Gust, B. Gonzalez-Morales, P. Gong, M.S. Wilbanks, I. Linkov, E.J. Perkins, A weight of evidence assessment approach for adverse outcome pathways, *Regul. Toxicol. Pharmacol.* 75 (2016) 46–57. doi:10.1016/j.yrtph.2015.12.014.
- [50] D.L. Filer, P. Kothiya, W.R. Setzer, R.S. Judson, M.T. Martin, The ToxCast™ Analysis Pipeline: An R Package for Processing and Modeling Chemical Screening Data, 2015. https://www.epa.gov/sites/production/files/2015-08/documents/pipeline_overview.pdf (accessed January 11, 2017).
- [51] U.S. EPA, ToxCast Chemical Inventory: Data Management and Data Quality Considerations, 2014. https://www.epa.gov/sites/production/files/2015-08/documents/toxcast_chemicals_qa_qc_management_141204.pdf (accessed January 13, 2017).
- [52] Leadscope, Leadscope, Inc, (2016). <http://www.leadscope.com/> (accessed March 23, 2017).

- [53] G. Roberts, G.J. Myatt, W.P. Johnson, K.P. Cross, P.E. Blower, LeadScope † : Software for Exploring Large Sets of Screening Data, *J. Chem. Inf. Comput. Sci.* 40 (2000) 1302–1314. doi:10.1021/ci0000631.
- [54] L.G. Valerio, C. Yang, K.B. Arvidson, N.L. Kruhlak, A structural feature-based computational approach for toxicology predictions, *Expert Opin. Drug Metab. Toxicol.* 6 (2010) 505–518. doi:10.1517/17425250903499286.
- [55] J.A. Cooper II, R. Saracci, P. Cole, Describing the validity of carcinogen screening tests, *Br. J. Cancer.* 39 (1979) 87–89.
- [56] OpenTox, Final database with additional content, (2011). <http://opentox.org/data/documents/development/opentoxreports/opentoxreportd34/view> (accessed October 14, 2016).
- [57] M. Gütlein, C. Helma, A. Karwath, S. Kramer, A Large-Scale Empirical Evaluation of Cross-Validation and External Test Set Validation in (Q)SAR, *Mol. Inform.* 32 (2013) 516–528. doi:10.1002/minf.201200134.
- [58] S.A. Rosenberg, M. Xia, R. Huang, N.G. Nikolov, E.B. Wedebye, M. Dybdahl, QSAR development and profiling of 72,524 REACH substances for PXR activation and CYP3A4 induction, *Comput. Toxicol.* (2017). doi:10.1016/j.comtox.2017.01.001.
- [59] ECHA, Substance information: tert-butyl-4-methoxyphenol, (2016). <https://echa.europa.eu/da/substance-information/-/substanceinfo/100.042.315>.
- [60] EFSA, Scientific opinion on the re-evaluation of butylated hydroxyanisole – BHA (E 320) as a food additive, 2011. doi:10.2903/j.efsa.2011.2392.
- [61] A. Pop, B. Kiss, F. Loghin, Endocrine disrupting effects of butylated hydroxyanisole (BHA - E320)., *Clujul Med.* 86 (2013) 16–20. <http://www.ncbi.nlm.nih.gov/pubmed/26527908> (accessed December 12, 2016).
- [62] NTP, Butylated Hydroxyanisole, (2016). <https://ntp.niehs.nih.gov/pubhealth/roc/index-1.html> (accessed March 21, 2017).
- [63] W. Mundy, S. Padilla, M. Gilbert, J. Breier, J. Cowden, K. Crofton, D. Herr, K. Jensen, K. Raffaele, N. Radio, K. Schumacher, Building a Database of Developmental Neurotoxicants: Evidence from Human and Animal Studies, *Toxicol.* 108. (2009). http://www.fluoridealert.org/wp-content/uploads/epa_mundy.pdf (accessed December 12, 2016).
- [64] C. V. Vorhees, R.E. Butcher, R.L. Brunner, V. Wootten, Developmental Neurobehavioral Toxicity of Butylated Hydroxyanisole (BHA) in Rats, *Neurobehav. Toxicol. Teratol.* 3 (1981) 321–329.
- [65] S.-H. Jeong, B.-Y. Kim, H.-G. Kang, H.-O. Ku, J.-H. Cho, Effects of butylated hydroxyanisole on the development and functions of reproductive system in rats, *Toxicology.* 208 (2005) 49–62. doi:10.1016/j.tox.2004.11.014.
- [66] S. Jobling, T. Reynolds, R. White, M.G. Parker, J.P. Sumpter, A variety of environmentally persistent chemicals, including some phthalate plasticizers, are weakly estrogenic, *Environ. Health Perspect.* 103 (1995) 582–587. doi:10.1289/ehp.95103582.
- [67] H.G. Kang, S.H. Jeong, J.H. Cho, D.G. Kim, J.M. Park, M.H. Cho, Evaluation of estrogenic and androgenic activity of butylated hydroxyanisole in immature female and castrated rats, *Toxicology.* 213 (2005) 147–156. doi:10.1016/j.tox.2005.05.027.
- [68] A.M. Soto, C. Sonnenschein, K.L. Chung, M.F. Fernandez, N. Olea, F.O. Serrano, The E-SCREEN Assay as a Tool to Identify Estrogens: An Update on Estrogenic Environmental Pollutants,

- Environ. Health Perspect. 103 (1995) 113–122. doi:10.1289/ehp.95103s7113.
- [69] M.G.R. ter Veld, B. Schouten, J. Louisse, D.S. van Es, P.T. van der Saag, I.M.C.M. Rietjens, A.J. Murk, Estrogenic Potency of Food-Packaging-Associated Plasticizers and Antioxidants As Detected in ER α and ER β Reporter Gene Cell Lines, *J. Agric. Food Chem.* 54 (2006) 4407–4416. doi:10.1021/jf052864f.
- [70] G. Würtzen, P. Olsen, BHA study in pigs, *Food Chem. Toxicol.* 24 (1986) 1229–1233. doi:10.1016/0278-6915(86)90311-X.
- [71] DK-EPA, List of Undesiable Substances 2009, (2009).
<http://www2.mst.dk/udgiv/publications/2011/05/978-87-92708-95-3.pdf> (accessed March 20, 2017).
- [72] DK-EPA, The EU list of potential endocrine disruptors, (2016).
<http://eng.mst.dk/topics/chemicals/endocrine-disruptors/the-eu-list-of-potential-endocrine-disruptors/> (accessed December 6, 2016).
- [73] U. Hass, S. Christiansen, M. Axelstad, J. Boberg, A. Andersson, N.E. Skakkebæk, K. Bay, H. Holbech, K.L. Kinnberg, P. Bjerregaard, Evaluation of 22 SIN List 2 . 0 substances according to the Danish proposal on criteria for endocrine disrupters, (2012) 1–141.
[http://eng.mst.dk/media/mst/67169/SIN report and Annex.pdf](http://eng.mst.dk/media/mst/67169/SIN_report_and_Annex.pdf) (accessed December 13, 2016).
- [74] SIN, SIN List result for CAS number 25013-16-6, (2016).
<http://sinlist.chemsec.org/search/search?query=25013-16-5> (accessed December 6, 2016).
- [75] EC, Commission Regulation (EU) 2015/282 of 20 February 2015 amending Annexes VIII, IX and X to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as regards the Extended One-Generation Reproductive Toxicity Study, (2015).
- [76] EFSA, OECD/EFSA Workshop on Developmental Neurotoxicity (DNT): the use of non-animal test methods for regulatory purposes, (2016).
<https://www.efsa.europa.eu/en/events/event/161018b> (accessed February 9, 2017).

3.2 QSAR Models for PXR Interaction and CYP3A4 Induction *In Vitro*

3.2.1 Published Paper

Computational Toxicology 1 (2017) 39–48



Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.elsevier.com/locate/comtox



QSAR development and profiling of 72,524 REACH substances for PXR activation and CYP3A4 induction



S.A. Rosenberg^a, M. Xia^b, R. Huang^b, N.G. Nikolov^{a,1}, E.B. Wedebye^{a,1}, M. Dybdahl^{a,*}

^a Division of Diet, Disease Prevention and Toxicology, National Food Institute, Technical University of Denmark, Mørkhøj Bygade 19, 2860 Søborg, Denmark

^b National Center for Advancing Translational Sciences, National Institutes of Health, 9800 Medical Center Drive, Rockville, MD 20850, USA

ARTICLE INFO

Article history:

Received 21 November 2016

Received in revised form 16 January 2017

Accepted 19 January 2017

Available online 24 January 2017

Keywords:

PXR
CYP3A4
QSAR
REACH
Screening

ABSTRACT

The Pregnane X Receptor (PXR) is a key regulator of enzymes, for example the cytochrome P450 isoform 3A4 (CYP3A4), and transporters involved in the metabolism and excretion of xenobiotics and endogenous compounds. Activation of PXR by xenobiotics causes altered protein expression leading to enhanced or decreased turnover of both xenobiotics and endogenous compounds. This can potentially result in perturbations of normal physiology and adverse effects. Identification of PXR activating and CYP3A4 inducing compounds is included in drug-discovery programs but we still need similar information for the remaining tens-of-thousands of man-made compounds to which humans are potentially exposed. In the present study, we used high-throughput *in vitro* assay results for 2816 drugs to develop four quantitative structure-activity relationship (QSAR) models with binary outputs for binding to the human PXR ligand binding domain, full-length human and rat PXR activation and human CYP3A4 induction, respectively. Rigorous cross- and blinded external validations demonstrated four robust and highly predictive models with balanced accuracies ranging from 75.4% to 92.7%. The models were applied to screen 72,524 substances pre-registered under the EU chemicals regulation, REACH, and the models could predict 52.5% to 71.9% of the substances within their respective applicability domains. These predictions can, for example, be used for priority setting and in weight-of-evidence assessments of chemicals. Statistical analyses of the experimental drug dataset and the QSAR-predicted set of REACH substances were performed to identify similarities and differences in frequencies of overlapping positive results for PXR binding, PXR activation and CYP3A4 induction between the two datasets.

© 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

The nuclear receptor (NR) superfamily is a large group of transcription factors that control expression of multiple genes involved

in a broad range of biological processes, such as development, homeostasis and metabolism. The transcriptional activity of NRs is primarily regulated through ligand binding [1]. The Pregnane X Receptor (PXR), first described by Kliewer and colleagues in 1998, is a member of the NR superfamily [2,3]. PXR is mainly expressed in the liver, intestine and kidneys, and plays a key role in the regulation of genes involved in the metabolism and efflux of endogenous hormones and xenobiotic molecules [3–5]. The genes regulated by PXR include genes encoding enzymes, such as cytochrome P450s (CYPs), glucuronyltransferases and sulfotransferases, as well as transporters, such as P-glycoprotein and multidrug resistance proteins [2,3,6–8]. The ligand-binding domain (LBD) of PXR is large and flexible, and can change its shape to accommodate structurally diverse molecules including steroids, bile acids, antibiotics, statins, and pesticides [9,10]. A considerable amount of inter-species variation has been observed in the PXR LBD with human, rabbit and rat sharing roughly 75–80% amino acid identity [11,12]. There are numerous examples of differences in ligand binding to PXR and resulting downstream transcription

Abbreviations: AD, applicability domain; AOP, adverse outcome pathway; CYP, cytochrome P450; CYP3A4, human cytochrome P450 isoform 3A4; DTU, Technical University of Denmark; Food, National Food Institute; hPXR, full-length human Pregnane X Receptor; hPXR-LBD, human Pregnane X Receptor Ligand Binding Domain; IATA, Integrated Approaches to Testing and Assessment; LBD, Ligand Binding Domain; LPDM, Leadscape® Predictive Data Miner; NCATS, National Center for Advancing Translational Sciences; NIH, National Institute of Health; NR, nuclear receptor; PLR, partial logistic regression; PRS, Pre-Registered Substances; PXR, Pregnane X Receptor; QSAR, quantitative structure-activity relationship; qHTS, quantitative high-throughput screening; REACH, Registration, Evaluation, Authorisation & restriction of Chemicals; rPXR, full-length rat Pregnane X Receptor; RXR α , Retinoid X Receptor α ; SD, standard deviation; TR-FRET, time-resolved fluorescence resonance energy transfer; XRE, Xenobiotic Response Element.

* Corresponding author.

E-mail address: mdyb@food.dtu.dk (M. Dybdahl).

¹ Contributed equally.

<http://dx.doi.org/10.1016/j.comtox.2017.01.001>

2468–1113/© 2017 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

of enzymes and transporters between species, which complicates the extrapolation of results from *in vivo* animal studies to humans [11,13–15].

PXR is located in the cytoplasm and translocated to the nucleus upon ligand binding, and here the PXR-ligand complex heterodimerizes with the Retinoid X Receptor alpha (RXR α), another member of the NR superfamily. The PXR-RXR α heterodimer complexes with co-activators, and this multi-protein complex binds to the Xenobiotic Response Element (XRE) in the promoter region of target genes and induces their transcription leading to altered expression of their encoded proteins [2,3,16]. Because many of the proteins regulated by PXR are not only involved in the metabolism and transport of xenobiotics, but also of various endogenous compounds such as steroid and thyroid hormones, an altered protein expression upon xenobiotic exposure may interfere with the homeostatic balance of such endogenous compounds [17,18]. This interference can potentially affect normal physiological functions [2,19] and may result in adverse health effects. Findings from previous studies indicate that there is an association between PXR activation by environmental chemicals and adverse health effects [15,18,20,21]. The importance of PXR activation is also reflected in a number of suggested adverse outcome pathways (AOPs) available from the online AOP-Wiki [22], for example an AOP describing how activation of PXR and other related NRs upregulate thyroid hormone catabolism resulting in hypothyroidism and subsequent adverse neurodevelopmental outcomes [23]. The AOPs are envisioned to promote the industry's and regulators' use of results from alternative methods such as *in vitro* tests and computational models in chemical risk assessments to reduce, refine or replace traditional animal tests [24–26], for example by applying the AOP in an Integrated Approaches to Testing Assessment (IATA) context to support regulatory decisions [27].

PXR is also known to be involved in drug-drug interactions in which an administered drug affects the metabolism and excretion of a co-administered drug, leading to decreased efficacy or increased toxicity [2,28,29]. For this reason, attenuation of PXR activity has become an important focus area in early drug-discovery programs [30]. Similar to drug-drug interactions, an altered expression of enzymes and transporters through PXR activation upon xenobiotic exposure may cause changes in the response to other xenobiotic compounds.

Among the many PXR target genes is the gene encoding CYP3A4, an oxidizing enzyme involved in phase I metabolism of various compounds [4,31]. CYP3A4 is considered the main drug-metabolizing CYP isoform in the human liver and is involved in the metabolism of more than 50% of drugs on the market [2,5]. In most cases, CYP3A4 causes chemicals to become less biologically active and promotes their excretion; but in other cases it has the opposite effect causing bioactivation by converting them to metabolites that are more toxic than the parent molecule [32].

Because xenobiotic activation of PXR has the potential to alter normal physiology and lead to adverse effects, it is of great importance to identify chemicals that may act through this mechanism. In a study from 2011, Shukla and colleagues used four high-throughput *in vitro* assays to profile more than 2800 clinically-used and investigational drugs for their ability to bind to the human PXR-LBD, activate full-length human and rat PXR, and induce human CYP3A4 [14]. Chemicals in the ToxCast program [33], which include both drugs and environmental chemicals, have also been tested for these mechanisms in related assays [34]. However, we still need similar information for the remaining tens-of-thousands of xenobiotics to which humans are potentially exposed [35,36].

In the present study, we used the high-throughput *in vitro* data from Shukla et al. [14] to train and validate four Quantitative Structure-Activity Relationship (QSAR) models for human PXR-

LBD binding, human and rat PXR activation, and human CYP3A4 induction, respectively. QSAR models are computational models that relate chemical structures to, e.g., a biological activity, and they can be used to predict the activity of an untested chemical based on its chemical structure (an introduction to QSAR can e.g. be found in [37,38]). In general, QSARs are rapid and cost-effective tools for predicting biological activities of chemical structures and can be used for virtual screening of single substances as well as large chemical inventories. The four developed models were applied to screen a structurally diverse library of 72,524 chemicals from the EU chemicals regulation REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) list of Pre-Registered Substances (PRS) [39,40], containing substances potentially present in our food, the environment and consumer products. These QSAR predictions can, e.g., be used, possibly together with other relevant data, 1) to identify and prioritize chemical substances for further testing and 2) in an IATA context, together with relevant AOP(s), to guide further testing and regulatory decisions in chemical risk assessments [25,27,41]. Furthermore, statistical analyses of the experimental drug dataset and the QSAR-predicted REACH PRS set were performed in order to elucidate similarities and differences in co-occurrences of overlapping positive results for PXR binding, PXR activation and CYP3A4 induction between the two chemical universes.

2. Materials and methods

2.1. Experimental datasets

We used four datasets containing chemical structure information and *in vitro* experimental data for a collection of 2816 clinically-used and investigational drugs to train and validate the QSAR models. The experimental data of the 2816 compounds included results from quantitative high-throughput screening (qHTS) for binding to the LBD of human PXR at the protein level (hPXR-LBD); activation of full-length human PXR (hPXR) and full-length rat PXR (rPXR) at the cellular level; and induction of human CYP3A4 at the cellular level (CYP3A4). All experimental data were generated by the National Center for Advancing Translational Sciences (NCATS) at the National Institute of Health (NIH). The compound collection, qHTS assays, and the classification of the qHTS results into actives, inconclusives and inactives have been described previously [14,42,43]. Briefly, actives showed binding to the hPXR-LBD, activation of hPXR and/or rPXR and/or induced transcription of CYP3A4 according to the applied assays. Inactives did not show activity in the given assay, and inconclusives showed equivocal activity results in the assays. Only the substances in each dataset classified as either active or inactive were used, i.e. substances with inconclusive experimental results were excluded. The experimental results for about one third of the substances in each of the four main datasets were masked by NIH NCATS and these compounds were used as external test sets for blinded external validations after the model development was finished. The selection of the test sets was designed and made by NIH NCATS scientists, who clustered all compounds in the dataset on structural similarity using the Euclidian distance and then, within each structure cluster and for each of the four endpoints, approximately one-third actives and one-third inactives were selected randomly. Thus the training and test sets are structurally comparable and have similar distributions of actives and inactives. NIH NCATS sent the training sets containing structure information and experimental results and the test sets containing only structure information to the National Food Institute (Food) at the Technical University of Denmark (DTU), who performed the structure preparations, the

model development and validations as well as the virtual screenings.

Furthermore, a dataset containing ~4000 additional compounds with experimental data from the qHTS assay for hPXR-LBD was used for supplementary performance assessment of the developed hPXR-LBD QSAR model [20,43].

2.2. Structural preparation of the datasets

The commercial QSAR software applied in this study can handle organic chemical substances with a known and unambiguous 2D structure. We apply an overall definition of substances acceptable for QSAR processing in all our in-house QSAR software [44,45], as substances:

- containing at least two carbon atoms
- containing only H, Li, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and/or I
- that are not mixtures containing two or more organic components

Substances that did not fulfil these criteria were removed from the datasets. Further processing of the structural information included dissociation simulation and subsequent neutralization of the structures, i.e. all substances were used in their non-ionized form. An overview of the number of QSAR-ready substances in the final training and external test sets after structure preparation can be found in Table 1. These sets are available upon request.

2.3. QSAR modeling

We used the commercial software, Leadscape® Predictive Data Miner (LPDM), a component of Leadscape® Enterprise Server version 3.2.4 [46], to build the four QSAR models. Briefly, LPDM calculates nine molecular descriptors (AlogP, Hydrogen Bond Acceptors and Donors, Lipinski Score, Molecular Weight, Parent Atom Number, Parent Molecular Weight, Polar Surface Area, Rotatable Bonds) for each chemical structure in the training set and performs a systematic sub-structural analysis using a template library of more than 27,000 predefined structural features [47]. The molecular descriptors and structural features are included in a default initial descriptor set. In addition, the system can generate and add training set-dependent structural features (scaffolds) to the descriptor set as well as remove redundant structural features from the descriptor set. Once a preliminary descriptor set has been created, an automatic descriptor selection procedure in LPDM selects the top 30% descriptors according to Yates X^2 -test for a binary response variable. A predictive model for a binary response variable is built using partial logistic regression (PLR) with further selection of descriptors in an iterative procedure, and selection of the optimum PLR factors based on least predictive residual sum of squares.

LPDM has the option of building composite binary models for training sets with a skewed distribution between the two activity classes, i.e. actives and inactives. With this option a number of sub-models are constructed, taking in each sub-model the entire smaller class, here the actives, and an equally large sample from the bigger class, here the inactives. The samples from the bigger class used in each of the sub-models are selected randomly but in such a way that their intersection is minimal and their union is the entire bigger class. The positive prediction probability (see Section 2.4) for a test chemical from a composite model is defined as the average of the positive prediction probabilities of all sub-models where the test chemical is in the structural domain [48]. Each sub-model in a composite model has its own unique set of selected descriptors and number of PLR factors.

We used five different modeling approaches in LPDM to build five predictive models for each of the four training sets: 1) single, 2) single with scaffolds, 3) single with scaffolds and reduced structural features, 4) composite, and 5) composite with scaffolds. In 1) and 4), the descriptors were selected among the default initial descriptor set, i.e. containing molecular descriptors and selected predefined structural features, and used to build a single model and a composite model, respectively. Next, scaffolds were generated in LPDM from the training set structures and added to the initial descriptor set, which subsequently was used for descriptor selection for models 2) and 5). In model 3), the scaffold-enriched descriptor set was reduced before descriptor selection by removing most similar structural features using a built-in function in LPDM. All models underwent a ten times two-fold cross-validation by the LPDM algorithm, which reuses the selected descriptor set from the parent model when building the cross-validation models [48]. For each of the four endpoints, we selected the predictive model with the highest performance from the LPDM cross-validation for further validation and screening studies (Fig. 1). The LPDM cross-validations were only applied for model selection and not used for model performance assessments. The four selected models were 'closed' for further development after this selection.

2.4. Applicability domain

Our definition of the applicability domain (AD) consists of two components: 1) the definition of a structural domain in LPDM, and 2) an in-house class probability refinement on the output from LPDM. For a test compound to be within LPDM's structural domain it is required that: all molecular descriptors used in the model can be calculated, it contains at least one structural feature used in the model, and that it has at least 30% Tanimoto similarity (default cut-off in the LPDM software) with a training set compound [48]. No prediction call (active/inactive) is generated by LPDM for a test compound outside this structural domain. For test compounds within the LPDM structural domain, a positive prediction probability, p , between 0 and 1, is given together with the prediction call; actives having a $p \geq 0.5$ and inactives having a $p < 0.5$ [48]. To

Table 1

Overview of the sizes of the training sets and the blinded external test sets used to develop and validate the four QSAR models. An extra dataset for hPXR-LBD binding was used for external validation. Substances with inconclusive experimental results were removed from the datasets.

Datasets	Training set			External test set		
	Total	Active (%)	Inactive (%)	Total	Active (%)	Inactive (%)
hPXR-LBD*	1537	143 (9.3)	1394 (90.7)	651	30 (4.6)	621 (95.4)
hPXR	1644	207 (12.6)	1437 (87.4)	702	59 (8.4)	643 (91.6)
rPXR	1671	97 (5.8)	1574 (94.2)	730	24 (3.3)	706 (96.7)
CYP3A4	1676	179 (10.7)	1497 (89.3)	715	45 (6.3)	670 (93.7)
Extra hPXR-LBD	–	–	–	2434	279 (11.5)	2155 (88.5)

* The experimental results of the test set were masked from the model developers at DTU Food by NIH NCATS until the models were developed and the test set had been predicted.

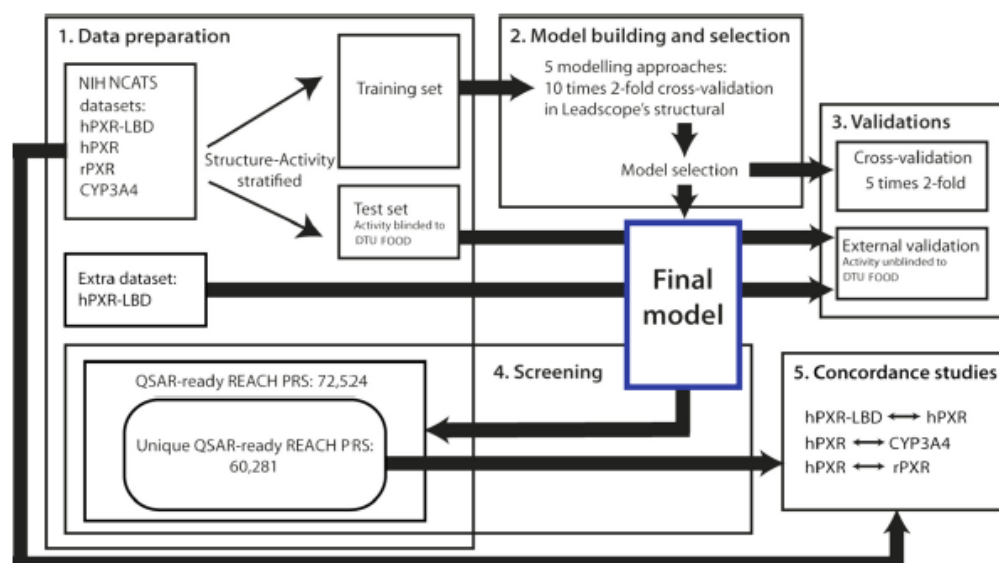


Fig. 1. Workflow of the modeling, screening and concordance rate studies.

exclude less reliable predictions, i.e. those with a positive prediction probability close to $p = 0.5$, we required $p \geq 0.7$ for active prediction calls and $p \leq 0.3$ for inactive prediction calls. Predictions within the LPDM structural domain but with an associated positive prediction probability in the interval 0.3 to 0.7 were defined as out of AD and excluded from the statistical analyses.

2.5. Cross- and external validation of the models

Each of the four selected predictive models was subsequently subject to a five times twofold stratified cross-validation procedure to estimate their robustness and predictive performance (Fig. 1). The applied procedure did not use the LPDM built-in cross-validation functionality. Instead, this was done by randomly removing 50% of the structures from the training set, keeping the ratio of actives and inactives. Then a cross-validation model was built from the reduced training set using the same modeling approach as in the parent model but by performing novel modeling where no information, such as selected descriptors, was reused from the parent model. The cross-validation model was applied to predict the removed 50%. Likewise, a cross-validation model was made on the removed 50% of the training set, and this model was used to predict the other 50%. This procedure was repeated five times resulting in ten cross-validation models. Sensitivity, specificity and balanced accuracy [49] were calculated for each of the ten cross-validation models, and from these the mean and standard deviation (SD) were computed to give an overall statistical estimate of the predictive performance and robustness of the full-training set parent model. Sensitivity is the percentage of experimental actives correctly predicted, specificity is the percentage of the experimental inactives correctly predicted, and balanced accuracy is the average of the sensitivity and specificity [49]. The coverage, i.e. the mean percentage of how many of the predicted substances that had predictions within the AD of the ten cross-validation models, was also calculated.

In addition, all four models underwent a blinded external validation using the experimentally masked test sets to further evaluate their predictive performance (Fig. 1). The prediction calls within the AD were compared to the experimental results, which

were made available to DTU Food by NIH NCATS after the model building step was finalized and the test sets predicted. The hPXR-LBD model underwent an additional external validation with the extra test set for hPXR-LBD. This external validation was not blinded, however, the data set was not applied in any of the model development or selection steps. Coverage, sensitivity, specificity and balanced accuracy were calculated for each model.

2.6. Screening of the REACH PRS inventory

The four selected and validated QSAR models were used to predict the activity of 72,524 substances from the REACH PRS list (Fig. 1). The REACH PRS chemical structures were extracted from the online Danish (Q)SAR Database structure set [44,45]. The structures were originally curated from deliverable 3.4 of the OpenTox EU project [39] and had been processed through the same structure preparation steps as described in Section 2.2 to meet the structural requirements from the QSAR modeling software. The proportion of the QSAR-predicted REACH PRS within the AD of each of the four models as well as the activity distributions of the predictions was calculated.

2.7. Concordance rates between endpoints

To study the co-occurrences in positive results for PXR binding, PXR activation and CYP3A4 induction, positive concordance rates both ways between the following endpoints were estimated:

- hPXR-LBD and hPXR,
- hPXR and rPXR, and
- hPXR and CYP3A4.

This was done for the full experimental drug datasets, i.e. the training and external test set data (excluding the extra hPXR-LBD test set) combined, as well as for the 60,281 unique structures out of the 72,524 QSAR-ready REACH PRS (Fig. 1).

For any endpoints, A and B, we used the following definition of the rate of actives in A also active in B, denoted *Concordance rate* ($A \rightarrow B$):

Concordance rate($A \rightarrow B$)

$$= \frac{\# \text{active in } A \text{ AND } B}{\# \text{active in } A \text{ AND } B + \# \text{active in } A \text{ AND in active in } B}$$

We apply the above definition twice for each pair of endpoints, A and B, to calculate Concordance rate ($A \rightarrow B$) and Concordance rate ($B \rightarrow A$).

For example, to assess the rate of hPXR-LBD ligands that activate hPXR, the following calculation was made:

Concordance rate(hPXR – LBD \rightarrow hPXR)

$$= \frac{\# \text{predicted/tested active in hPXR-LBD AND hPXR}}{\# \text{predicted/tested active in hPXR-LBD AND hPXR} + \# \text{predicted/tested active in hPXR – LBD AND in active in hPXR}}$$

Likewise, the concordance rate for hPXR activators that were also active for binding to hPXR-LBD was calculated as:

Concordance rate(hPXR \rightarrow hPXR – LBD)

$$= \frac{\# \text{predicted/tested active in hPXR-LBD AND hPXR}}{\# \text{predicted/tested active in hPXR – LBD AND hPXR} + \# \text{predicted/tested active in hPXR AND in active in hPXR-LBD}}$$

Differences and similarities between corresponding concordance rates in the drug and REACH PRS universes were identified.

3. Results

3.1 Predictive performance and robustness

For each of the four endpoints the model with the highest performance from the LPDM cross-validation was selected for further validation and screening studies. The four selected models were all composite models consisting of seven to ten sub-models. Each of the four selected models underwent both an in-house rigorous five times leave-50%-out cross-validation and a DTU Food blinded external validation to assess their robustness and predictive performance within the defined AD. The validation results are presented in Table 2 together with information about the number of sub-models in the selected composite model. Overall, the results presented in Table 2 show that the rigorous leave-50%-out cross-validations underestimated the models' predictive performances compared to the blinded external validations. The models will be made available for prediction of user-submitted structures in a coming free online Danish (Q)SAR Models sister-site to the Danish (Q)SAR database at the DTU homepage [45].

3.2. Screening of the REACH PRS inventory

A set of 72,524 substances from the REACH PRS list was screened through the four QSAR models. Of the 72,524 REACH PRS, 28.6% (20,727) were in the common AD of all four models, and of these, 1.5% corresponding to 320 substances were predicted active for all four endpoints and 77.1% corresponding to 15,979 substances were predicted inactive by all four models. The number of REACH PRS predicted within the defined AD of each model and the distribution of active and inactive predictions are given in Table 3.

3.3. Concordance rates between hPXR-LBD binding and Full-Length hPXR activation

The cell-free hPXR-LBD assay is a LanthaScreen TR-FRET (time resolved fluorescence resonance energy transfer)-based assay that identifies binding of a chemical to the LBD of human PXR, whereas the cell-based hPXR assay identifies compounds that can activate human full-length PXR either through direct LBD binding or through other signaling pathways [50,51]. In order to obtain more

information on frequencies of possible mechanisms of PXR activation for drugs and REACH PRS, we calculated two-way concordance

rates between hPXR-LBD binding and full-length hPXR activation for the experimental results of the full drug datasets and for the QSAR predictions of the REACH sets, respectively (Fig. 2a). For the experimental drug data the rate of hPXR-LBD tested binders resulting in hPXR activation was 44.0% (63/(63 + 79)), and the rate of hPXR activators binding to hPXR-LBD was 37.7% (63/(63 + 104)). For the predicted REACH substances only compounds in the common AD of the two models ($n = 22,486$) were included in the analysis, and among these 2624 were predicted active by both models and 16,842 were predicted inactive by both models. Of the remaining 3020 discordant predictions, 2408 were predicted active for hPXR-LBD but inactive for hPXR, while 612 were predicted active for hPXR but inactive for hPXR-LBD. Based on these predictions, it was estimated that 52.1% (2624/(2,624 + 2408)) of the predicted hPXR-LBD actives are also predicted to cause hPXR activation, whereas 81.1% (2624/(2,624 + 612)) of the predicted hPXR activators are also predicted to bind to hPXR-LBD.

3.4. Concordance rates between hPXR activation and CYP3A4 induction

Since PXR is known to induce the transcription of CYP3A4 [4,31], we calculated the concordance rates between hPXR activation and CYP3A4 induction for both the tested drugs and the QSAR-predicted REACH substances set (Fig. 2b). For the experimental drug data, the rate of hPXR active drugs that result in CYP3A4 induction was 53.6% (113/(113 + 98)), and the rate of CYP3A4 inducers also activating hPXR was 66.5% (113/(113 + 57)). Of the 24,364 REACH PRS predicted within the common AD of the two models, 2945 were predicted active by both models, whereas 20,960 were predicted inactive in both models. Among the 459 substances with discrepant predictions, 385 were predicted active by hPXR only and 74 were predicted active only by the CYP3A4 model. From these numbers it can be estimated that 88.4% (2945/(2945 + 459)) of the REACH substances predicted to cause hPXR activation were also predicted to induce CYP3A4, and that 97.5% (2945/(2945 + 74)) of the predicted CYP3A4 inducing REACH substances were also predicted to activate hPXR.

Table 2

Coverage and predictive performance of the four QSAR models. Only predictions inside the defined AD were included in the statistical analyses.

QSAR model	Statistical parameter	Cross-validation,% (SD,%) 5 times 2-fold*	External validation,% (actual numbers)	
			Blinded test sets**	Extra hPXR-LBD test set
hPXR-LBD Approach 5) 10 sub-models	Coverage	66.0 (3.3)	67.3 (438/651)	60.6 (1475/2434)
	Sensitivity	68.7 (7.3)	85.0 (17/20)	71.9 (97/135)
	Specificity	84.5 (2.0)	87.8 (367/418)	80.4 (1078/1340)
	Balanced accuracy	76.6 (3.2)	86.4	76.1
hPXR Approach 5) 7 sub-models	Coverage	60.3 (2.9)	59.1 (415/702)	–
	Sensitivity	72.5 (6.7)	80.0 (24/30)	–
	Specificity	80.4 (3.7)	85.2 (328/385)	–
	Balanced accuracy	76.4 (2.9)	82.6	–
rPXR Approach 4) 10 sub-models	Coverage	74.0 (3.0)	80.0 (584/730)	–
	Sensitivity	58.9 (11.0)	91.3 (21/23)	–
	Specificity	92.0 (2.4)	94.1 (528/561)	–
	Balanced accuracy	75.4 (4.7)	92.7	–
CYP3A4 Approach 5) 9 sub-models	Coverage	64.7 (3.0)	63.4 (453/715)	–
	Sensitivity	71.6 (7.6)	76.9 (20/26)	–
	Specificity	80.7 (2.7)	85.5 (365/427)	–
	Balanced accuracy	76.1 (3.3)	81.2	–

* A five times twofold cross-validation with same active-inactive ratio as the full training set and without reusing selected descriptors from the parent model. Coverage, sensitivity and specificity are the mean from the ten cross-validation models with the standard deviation (SD) in parentheses.

** The experimental results of the test set structures were made available to DTU Food by NIH NCATS after they had been predicted in the respective models by DTU Food.

Table 3

Prediction and domain results for the 72,524 REACH PRS.

QSAR model	Total in AD (%)	Predicted Active in AD (%)	Predicted Inactive in AD (%)
hPXR-LBD	43,551 (60.1)	11,490 (26.4)	32,061 (73.6)
hPXR	38,114 (52.5)	6167 (16.2)	31,947 (83.8)
rPXR	52,144 (71.9)	3141 (6.0)	49,003 (94.0)
CYP3A4	42,861 (59.1)	5874 (13.7)	36,987 (86.3)

3.5. Concordance rates between human and rat Full-Length PXR activation

Species differences in PXR activation by chemicals have previously been identified [11,14,52] and information on these differences can be of importance when extrapolating data from rat *in vivo* studies to humans, e.g. in chemical risk assessment. In the experimental drug dataset, the rate of human PXR activating drugs that also activate the rat PXR was 25.9% (51/(51 + 146)) (Fig. 2c). Conversely, 56.7% (51/(51 + 39)) of the rat PXR activating drugs also activated human PXR. To estimate the species differences in human and rat PXR activation with regard to the QSAR-predicted REACH substances, we compared REACH PRS QSAR-predictions from the hPXR and rPXR models. Among the 25,498 REACH PRS predicted in the common AD, 862 were predicted active in both models, 2788 were predicted active for hPXR only, and 573 were predicted active for rPXR only. The remaining 21,275 were predicted inactive by both models. From this it can be estimated that 23.6% (862/(862 + 2788)) of the QSAR-predicted REACH PRS activating human PXR were also predicted to activate rat PXR, and 60.1% (862/(862 + 573)) of the predicted rat PXR activators were also predicted as human PXR activators.

4. Discussion

In the present study, we developed four global binary QSAR models for human PXR-LBD binding, human and rat full-length PXR activation, and human CYP3A4 induction, respectively. The models were used to screen more than 70,000 REACH substances. To our knowledge this is the first study to profile a large set of chemical substances potentially used in industrial processes, food and consumer products, such as cleaning products, paints, clothes,

and furniture, by QSAR with respect to both PXR binding/activation and CYP3A4 induction.

4.1. Predictive performance and robustness

A number of different modeling approaches in LPDM were used to build models on the four training sets and the best performing model for each endpoint was selected for further validation studies and screening of the REACH PRS inventory. It is known that sensitivity and specificity of binary models can, depending on the applied modeling algorithm, be affected by the distribution of actives and inactives in the training set. A training set with a greater number of inactives will often result in a higher specificity at the expense of sensitivity and vice versa in the case of overrepresentation of actives. This is likely the reason why the single models built with the full, imbalanced training sets were outperformed by the composite models: all four selected models were composite models consisting of seven to ten sub-models with balanced sub-training sets. The composite model feature in LPDM was implemented to handle imbalanced training sets [48], in this case training sets with only 5.8% to 12.6% actives.

All four models showed high predictive performances with balanced accuracies in the external validations ranging from 76.1% to 92.7% (Table 2). Both the high quality of the experimental data originating from robust assays [14,53] as well as the composite modeling approach in LPDM have undoubtedly contributed to the high performances of the models. The cross-validation results were generally pessimistic compared to the external validations (Table 2), especially with regard to the sensitivity. The fact that the cross-validation results in this study are pessimistic compared to the external validations is in accordance with the finding in, e.g. [54], where this issue was systematically studied. The generally low standard deviations (SDs) in the cross-validations indicate robust models, i.e. their performances are not drastically altered in response to perturbations of the training set composition. Both the remarkably lower cross-validation sensitivities relative to the external validation sensitivities and their higher SDs is likely due to the rigorous cross-validation procedure of removing 50% of the few actives in the non-congeneric training sets. The effects of removing 50% is most clearly reflected in the rPXR model (Table 2), which was also the model with the fewest training set actives, i.e. 97 actives (Table 1). Often *k*-fold cross-validations of models built from training sets of similar size as those in this study are

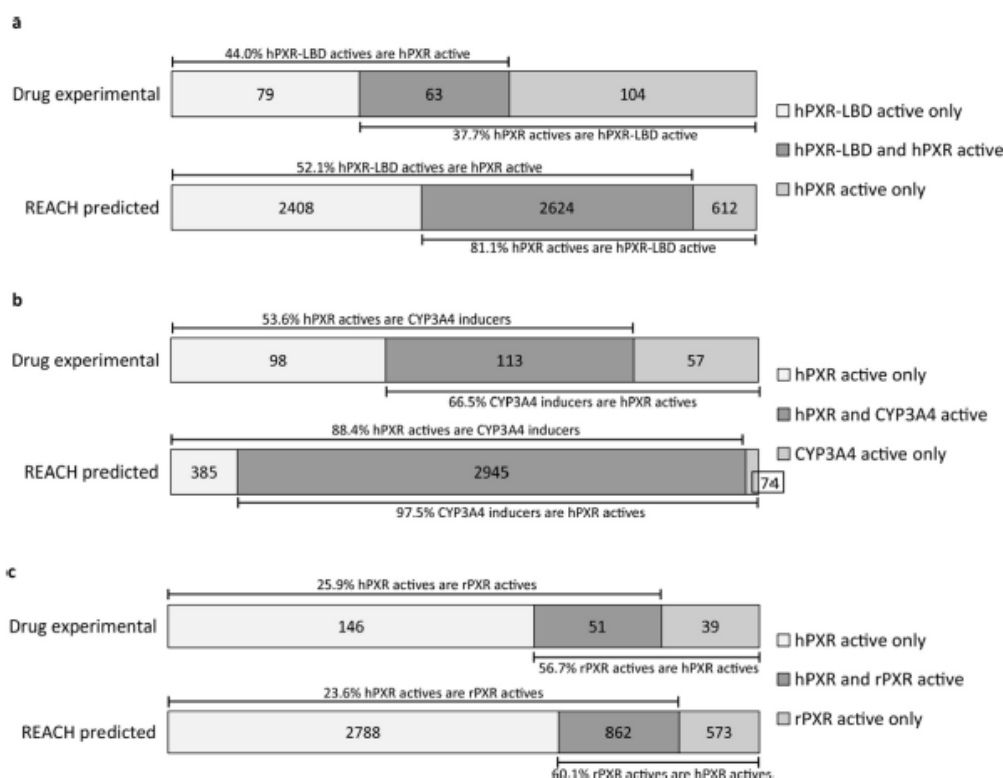


Fig. 2. Overlap of positive results between two endpoints and two-way concordance rates. a) comparing tested/predicted hPXR-LBD binders with hPXR activators, b) comparing tested/predicted hPXR activators with CYP3A4 inducers, and c) comparing tested/predicted rPXR activators with hPXR activators.

performed by removing 10% or 20% (i.e., $k = 10$ or 5) of the training set, leaving more data to train the cross-validation models [52,54]. The cross-validation results indicate that the leave-50%-out cross-validation performed in the present study was causing too big perturbations. Retrospectively, it seems that a 10 or 20%-leave-out cross-validation would have been more appropriate in this case.

The hPXR-LBD model in the present study has a lower cross-validation sensitivity (68.7%) compared to a similar hPXR-LBD model from Dybdahl and colleagues (82.3%) [14,20,43]. The difference in sensitivities is likely due to differences in the composition of the two training sets, with the Dybdahl model having more than twice as many actives in its training set, i.e. 299 versus 143 actives in the current model, leaving more actives for the 50% reduced cross-validation models. Additionally, the Dybdahl model cross-validation [20] was performed using LPDMs algorithm, which, we have experienced in some cases, returns overoptimistic statistics because of its reuse of parent model descriptors in the cross-validation models.

The size of the DTU Food masked external test sets with predictions inside the respective model's AD ranged from 415 to 584 structures, with 20–30 structures having active experimental results (Table 2). In general, external test sets should be sufficiently large and representative of the model's AD to ensure that the predictive performance results are not random. The distributions of experimentally active and inactive structures in these external test sets are imbalanced toward more inactives similar to the training set distributions. Although the masked test sets in total are quite large for external validation, the few actives make the calculations of sensitivity less robust. The supplementary external validation of the hPXR-LBD model included 135 experimentally active sub-

stances out of the total 1475 test set structures predicted inside the hPXR-LBD model's AD (Table 2). This larger number of actives may provide a more accurate estimate of the hPXR-LBD model's sensitivity compared to the result from the blinded external validation with only 20 experimentally active compounds. The extra external validation of the hPXR-LBD model resulted in overall lower predictive performance estimates compared to the blinded external validation (Table 2). This can be due to differences in the chemical universes of the two test set with the blinded test set likely representing the training set better due to the chemical-similarity test set selection procedure described in Section 2.1 [55,56]. A previous study have shown that this type of rational test set selection can give optimistic validation results [57]. Also, although the hPXR-LBD data in the two datasets were generated using the same assay protocol in the same laboratory, minor differences in the data analysis of the extra hPXR-LBD dataset compared to that of the NIH NCATS hPXR-LBD data could have negatively affected the validation results to some degree. Available ToxCast datasets [34] with experimental results for human PXR binding and activation and CYP3A4 induction were not applied in the validation study due, in our opinion, to large dissimilarities in the assay protocols and data analysis with the NIH NCATS training sets.

4.2. Screening of the REACH PRS inventory

The four selected models were used to predict 72,524 REACH PRS in order to give an estimate of the number of PXR activators and CYP3A4 inducers in this chemical universe (Table 3). A large overlap in the chemical similarity of small molecule drugs and

environmental chemicals has been identified, and other QSAR models trained on drug data have been shown to have a high predictability of environmental chemicals [52,55,56]. This, together with the application of a structural AD to avoid extrapolations, justifies the use of the drug-data trained models to screen the REACH set. The screening indicates that the predicted REACH PRS set contains nearly the same rate of human and rat PXR full-length activators as well as CYP3A4 inducers compared to the experimentally tested drugs in the training sets, i.e. 16% vs. 13%, 6% vs. 5.8%, and 14% vs. 11%, respectively. The hPXR-LBD model, however, predicted 26% of the REACH PRS inside the model's AD to be hPXR-LBD ligands, which was remarkably higher than the 9.3% hPXR-LBD active drugs in the training set. Since the hPXR-LBD model does not seem to be biased towards producing many false positive predictions based on the high specificity in the three validations, i.e. 80.4% to 87.8% (Table 2), this is unlikely the only reason for the high prevalence of predicted hPXR-LBD actives in the REACH PRS set. The increased focus on attenuation of PXR activity and the introduction of a filtering procedure in early drug development [30] might to some degree explain the nearly three-fold lower rate of hPXR-LBD ligands among drugs compared to the predicted REACH substances.

4.3. Concordance rates between endpoints

The calculated concordance rates between endpoints using either experimental test results or QSAR predictions can provide information on the frequencies of the possible mechanisms by which chemicals act as well as reveal differences and similarities between the two chemical inventories (Fig. 2). Results from a previous study indicate that differences in the biological mechanisms of drugs and environmental chemicals exist [58]. When concordance rates are based on QSAR predictions, they can be influenced by the uncertainty inherent in the predictive models, but since all four models had high predictive performances in the external validations (Table 2), we expect this uncertainty to be fairly low. For the concordance rates based on the experimental data, these can be affected by the fact that experimental tests may not be 100% reproducible. In a follow-up study, Shukla and colleagues [14] retested 72 compounds in the four qHTS assays and the activities were confirmed for 71 (hPXR-LBD), 66 (hPXR), 72 (rPXR) and 70 (CYP3A4) of the compounds, respectively, with no information of the activity distribution. This could indicate a slightly higher rate of false positive and/or false negative test results in the hPXR assay compared to the other three assays. Inclusion of false positives and/or negatives in the hPXR experimental data could in this case have affected the hPXR model development and its performance measurements as well as the subsequent concordance rate studies of both the experimental and predicted datasets.

Roughly half of the hPXR-LBD binders were also hPXR activators for both the tested drugs (44%) and the predicted REACH PRS (52%) (Fig. 2a). This may reflect that the ~50% active compounds from the hPXR-LBD cell-free assay that are not active in the cell-based hPXR activation assay either cannot enter the cell, are biodegraded in the cellular environment, or act as human PXR antagonists [14,28]. For the hPXR activators that were also hPXR-LBD ligands, we observed a difference in the concordance rates between the two universes, with only 38% of the full-length hPXR activators being hPXR-LBD ligands for the tested drugs as opposed to 81% for the QSAR-predicted REACH PRS. This difference might be a reflection of the approximately three-times higher occurrence of predicted hPXR-LBD binders in the QSAR-predicted REACH PRS universe and thus a higher chance for hPXR activators to also be predicted active by hPXR-LBD. The part of the hPXR activators that were not hPXR-LBD ligands likely exert their effect on PXR activation through other signaling pathways such as protein kinase path-

ways [50,51]. They may also be chemicals that are not able to displace the tracer molecule in the hPXR-LBD assay [14], a known problem with LanthaScreen TR-FRET-based binding assays.

When comparing hPXR activation and CYP3A4 induction higher concordance rates were found for the QSAR-predicted REACH PRS than for the tested drugs (Fig. 2b). Among the REACH PRS predictions, 88.4% of the hPXR activators also induced CYP3A4, while for the experimentally tested drugs this was only the case for 53.5% of the hPXR activators. Multiple factors can explain the absence of CYP3A4 induction by hPXR activators, for example, negative feedback loops repressing CYP3A4 expression, differences in recruitment of co-activators resulting in variations in the promoter region binding and downstream gene transcription patterns [59], as well as assay-related biochemical limitations [60]. Of the CYP3A4 inducers, 97.5% and 66.5% of the predicted REACH PRS and tested drugs, respectively, were also hPXR activators. An explanation to why some CYP3A4 inducers were not hPXR activators could be that other transcription factors or signaling pathways in the cell have led to the CYP3A4 induction. The high concordance rates of 97.5% and 88.4% between the prediction sets indicate that the two models have high agreement in their predictions.

Previous studies have reported species differences between human and rat PXR ligands [14,15,52,61] and this is supported by a highly divergent inter-species PXR-LBD amino acid sequence [11] with human and rat PXR-LBD sharing only 78.3% amino acid sequence similarity according to a calculation made using the web-based SeqAPASS software [62]. In the present study, around 25% of the hPXR activators among both the tested drugs and the predicted REACH PRS were also activating rPXR (Fig. 2c). Among the rPXR activators 57–60% in both universes were also activating hPXR. These results support that species differences in chemical action of drugs and REACH substances on PXR exist. The current study has identified 3361 (2788 + 573) REACH substances for which extra attention is necessary when extrapolating rat *in vivo* data to humans.

Overall, this statistical analysis indicates that QSAR predictions of larger chemical inventories can be applied to study overlap in activities between biological endpoints. Such studies can potentially be used in hypotheses generation of new mechanistic associations.

5. Conclusions

We have developed four QSAR models for human PXR-LBD binding, human and rat full-length PXR activation, and human CYP3A4 induction. All four models were robust with high predictive performances. The models were used to screen a set of 72,524 REACH PRS and of the QSAR-predicted REACH substances the number of actives were as follows: hPXR-LBD (11,490), hPXR (6167), rPXR (3141), and CYP3A4 (5874). Furthermore, the experimental data and the predictions of the REACH substances were analyzed to obtain information on co-occurrences of positive results for PXR activation and CYP3A4 induction in the two chemical universes. The developed models can in a fast and cost-efficient way provide information that can be used for prioritization purposes as well as in combination with other data in IATAs including weight-of-evidence assessments of chemical substances. The models can also help in future design of safer chemicals and drugs.

Conflict of interest statement

The authors declare that they have no conflict of interest in relation with this paper.

Acknowledgements

We would like to thank the Danish 3R Center and the Danish Environmental Protection Agency for supporting the project.

References

- [1] D.J. Mangelsdorf, C. Thummel, M. Beato, P. Herrlich, G. Schütz, K. Umesono, B. Blumberg, P. Kastner, M. Mark, P. Chambon, R.M. Evans, The nuclear receptor superfamily: the second decade, *Cell* 83 (1995) 835–839, [http://dx.doi.org/10.1016/0092-8674\(95\)90199-X](http://dx.doi.org/10.1016/0092-8674(95)90199-X).
- [2] A. di Masi, E. De Marinis, P. Ascenzi, M. Marino, Nuclear receptors CAR and PXR: molecular, functional, and biomedical aspects, *Mol. Aspects Med.* 30 (2009) 297–343, <http://dx.doi.org/10.1016/j.mam.2009.04.002>.
- [3] S.A. Klierer, J.T. Moore, L. Wade, J.L. Staudinger, M.A. Watson, S.A. Jones, D.D. McKee, B.B. Oliver, T.M. Willson, R.H. Zetterstrom, T. Perlmann, J.M. Lehmann, An orphan nuclear receptor activated by pregnanes defines a novel steroid signaling pathway, *Cell* 92 (1998) 73–82, [http://dx.doi.org/10.1016/S0092-8674\(00\)80900-9](http://dx.doi.org/10.1016/S0092-8674(00)80900-9).
- [4] G. Bertilsson, J. Heidrich, K. Svensson, M. Åsman, L. Jendeberg, M. Sydow-Backman, R. Ohlsson, H. Postlind, P. Blomquist, A. Berkenstam, Identification of a human nuclear receptor defines a new signaling pathway for CYP3A induction, *Proc. Natl. Acad. Sci.* 95 (1998) 12208–12213, <http://dx.doi.org/10.1073/pnas.95.21.12208>.
- [5] J.M. Lehmann, D.D. McKee, M.A. Watson, T.M. Willson, J.T. Moore, S.A. Klierer, The human orphan nuclear receptor PXR is activated by compounds that regulate CYP3A4 gene expression and cause drug interactions, *J. Clin. Invest.* 102 (1998) 1016–1023, <http://dx.doi.org/10.1172/JCI3703>.
- [6] A.H. Tolson, H. Wang, Regulation of drug-metabolizing enzymes by xenobiotic receptors: PXR and CAR, *Adv. Drug Deliv. Rev.* 62 (2010) 1238–1249, <http://dx.doi.org/10.1016/j.addr.2010.08.006>.
- [7] C. Xu, C.Y.-T. Li, A.-N.T. Kong, Induction of phase I, II and III drug metabolism/transport by xenobiotics, *Arch. Pharm. Res.* 28 (2005) 249–268, <http://dx.doi.org/10.1007/BF02977789>.
- [8] D. Gardner-Stephen, J.-M. Heydel, A. Goyal, Y. Lu, W. Xie, T. Lindblom, P. Mackenzie, A. Radominska-Pandya, Human PXR variants and their differential effects on the regulation of human UDP-glucuronyltransferase gene expression, *Drug Metab. Dispos.* 32 (2004) 340–347, <http://dx.doi.org/10.1124/dmd.32.3.340>.
- [9] R.E. Watkins, G.B. Wisely, L.B. Moore, J.L. Collins, M.H. Lambert, S.P. Williams, T. M. Willson, S.A. Klierer, M.R. Redinbo, The human nuclear xenobiotic receptor PXR: structural determinants of directed promiscuity, *Science* (80-) 292 (2001) 2329–2333, <http://dx.doi.org/10.1126/science.1060762>.
- [10] V. Delfosse, B. Dendele, T. Huet, M. Grimaldi, A. Boulahtouf, S. Gerbal-Chaloin, B. Beucher, D. Roelckin, C. Muller, R. Rahmani, V. Cavallès, M. Daujat-Chavanieu, V. Vivat, J.M. Pascussi, P. Balaguer, W. Bourguet, Synergistic activation of human pregnane X receptor by binary cocktails of pharmaceutical and environmental compounds, *Nat. Commun.* 6 (2015) 1–10, <http://dx.doi.org/10.1038/ncomms9089>.
- [11] S.A. Jones, L.B. Moore, J.L. Shenk, G.B. Wisely, G.A. Hamilton, D.D. McKee, N.C.O. Tomkinson, E.L. LeCluyse, M.H. Lambert, T.M. Willson, S.A. Klierer, J.T. Moore, The pregnane X receptor: a promiscuous xenobiotic receptor that has diverged during evolution, *Mol. Endocrinol.* 14 (2000) 27–39, <http://dx.doi.org/10.1210/med.14.1.0409>.
- [12] H. Zhang, E. LeCluyse, L. Liu, M. Hu, L. Matoney, W. Zhu, B. Yan, Rat pregnane X receptor: molecular cloning, tissue distribution, and xenobiotic regulation, *Arch. Biochem. Biophys.* 368 (1999) 14–22, <http://dx.doi.org/10.1006/abbi.1999.1307>.
- [13] E.L. LeCluyse, Pregnane X receptor: molecular basis for species differences in CYP3A induction by xenobiotics, *Chem. Biol. Interact.* 134 (2001) 283–289, [http://dx.doi.org/10.1016/S0009-2797\(01\)00163-6](http://dx.doi.org/10.1016/S0009-2797(01)00163-6).
- [14] S.J. Shukla, S. Sakamuru, R. Huang, T.A. Moeller, P. Shinn, D. VanLeer, D.S. Auld, C.P. Austin, M. Xia, Identification of clinically used drugs that activate pregnane X receptors, *Drug Metab. Dispos.* 39 (2011) 151–159, <http://dx.doi.org/10.1124/dmd.110.035105>.
- [15] Y. Sui, S.-H. Park, R.N. Helsley, M. Sunkara, F.J. Gonzalez, A.J. Morris, C. Zhou, Bisphenol A increases atherosclerosis in pregnane X receptor-humanized ApoE deficient mice, *J. Am. Hear. Assoc.* 3 (2014) 1–11, <http://dx.doi.org/10.1161/JAHA.113.000492>.
- [16] E.J. Squires, T. Sueyoshi, M. Negishi, Cytoplasmic localization of pregnane X receptor and ligand-dependent nuclear translocation in mouse liver, *J. Biol. Chem.* 279 (2004) 49307–49314, <http://dx.doi.org/10.1074/jbc.M407281200>.
- [17] M.N. Jacobs, G.T. Nolan, S.R. Hood, Lignans, bacteriocides and organochlorine compounds activate the human pregnane X receptor (PXR), *Toxicol. Appl. Pharmacol.* 209 (2005) 123–133, <http://dx.doi.org/10.1016/j.taap.2005.03.015>.
- [18] X.C. Kretschmer, W.S. Baldwin, CAR and PXR: xenosensors of endocrine disruptors?, *Chem. Biol. Interact.* 155 (2005) 111–128, <http://dx.doi.org/10.1016/j.cbi.2005.06.003>.
- [19] N.K. Chaturvedi, S. Kumar, S. Negi, R.K. Tyagi, Endocrine disruptors provoke differential modulatory responses on androgen receptor and pregnane and xenobiotic receptor: potential implications in metabolic disorders, *Mol. Cell. Biochem.* 345 (2010) 291–308, <http://dx.doi.org/10.1007/s11010-010-0583-6>.
- [20] M. Dybdahl, N.G. Nikolov, E.B. Wedebye, S.O. Jónsdóttir, J.R. Niemelä, QSAR model for human pregnane X receptor (PXR) binding: screening of environmental chemicals and correlations with genotoxicity, endocrine disruption and teratogenicity, *Toxicol. Appl. Pharmacol.* 262 (2012) 301–309, <http://dx.doi.org/10.1016/j.taap.2012.05.008>.
- [21] I. Shah, K. Houck, R.S. Judson, R.J. Kavlock, M.T. Martin, D.M. Reif, D.J. Dix, Using nuclear receptor activity to stratify hepatocarcinogens, *PLoS ONE* 6 (2011) e14584, <http://dx.doi.org/10.1371/journal.pone.0014584>.
- [22] AOP-wiki, AOP-Wiki homepage, 2016. https://aopwiki.org/wiki/index.php/Main_Page (accessed October 6, 2016).
- [23] AOP8, Aop:8 – Upregulation of Thyroid Hormone Catabolism via Activation of Hepatic Nuclear Receptors, and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, 2016. <https://aopwiki.org/wiki/index.php/Aop:8> (accessed October 6, 2016).
- [24] G.T. Ankley, R.S. Bennett, R.J. Erickson, D.J. Hoff, M.W. Homung, R.D. Johnson, D.R. Mount, J.W. Nichols, C.L. Russom, P.K. Schmieder, J.A. Serrano, J.E. Tietje, D.L. Villeneuve, Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment, *Environ. Toxicol. Chem.* 29 (2010) 730–741, <http://dx.doi.org/10.1002/etc.34>.
- [25] S. Gutsell, P. Russell, The role of chemistry in developing understanding of adverse outcome pathways and their application in risk assessment, *Toxicol. Res. (Camb.)* 2 (2013) 299, <http://dx.doi.org/10.1039/c3tx50024a>.
- [26] N.C. Meinstreuer, K. Sullivan, D. Allen, S. Edwards, D.L. Mendrick, M. Embry, J. Matheson, J.C. Rowlands, S. Munn, E. Maull, W. Casey, Adverse outcome pathways: from research to regulation scientific workshop report, *Regul. Toxicol. Pharmacol.* 76 (2016) 39–50, <http://dx.doi.org/10.1016/j.yrtph.2016.01.007>.
- [27] K.E. Tollefsen, S. Scholz, M.T. Cronin, S.W. Edwards, J. de Knecht, K. Crofton, N. Garcia-Reyero, T. Hartung, A. Worth, G. Patlewicz, Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA), *Regul. Toxicol. Pharmacol.* 70 (2014) 629–640, <http://dx.doi.org/10.1016/j.yrtph.2014.09.009>.
- [28] S. Ekins, C. Chang, S. Mani, M.D. Krasowski, E.J. Reschly, M. Iyer, V. Kholodovych, N. Ai, W.J. Welsh, M. Sinz, P.W. Swaan, R. Patel, K. Bachmann, Human pregnane X receptor antagonists and agonists define molecular requirements for different binding sites, *Mol. Pharmacol.* 72 (2007) 592–603, <http://dx.doi.org/10.1124/mol.107.038398>.
- [29] E. Qiao, M. Ji, J. Wu, R. Ma, X. Zhang, Y. He, Q. Zha, X. Song, L.-W. Zhu, J. Tang, Expression of the PXR gene in various types of cancer and drug resistance (Review), *Oncol. Lett.* 5 (2013) 1093–1100, <http://dx.doi.org/10.3892/ol.2013.1149>.
- [30] Y. Gao, S.H. Olson, J.M. Balkovec, Y. Zhu, I. Royo, J. Yabut, R. Evers, W. Tang, D.P. Hartley, R.T. Mosley, S.H. Olson, J.M. Balkovec, Y. Zhu, I. Royo, J. Yabut, R. Evers, E.Y. Tan, W. Tang, D.P. Hartley, R.T. Mosley, Attenuating pregnane X receptor (PXR) activation: a molecular modelling approach, *Xenobiotica* 37 (2007) 124–138, <http://dx.doi.org/10.1080/00498250601050412>.
- [31] M. Sinz, G. Wallace, J. Sahi, Current industrial practices in assessing CYP450 enzyme induction: preclinical and clinical, *AAPS* 10 (2008) 391–400, <http://dx.doi.org/10.1208/s12248-008-9037-4>.
- [32] J.E. Laine, S. Auriola, M. Pasanen, R.O. Juvonen, Acetaminophen bioactivation by human cytochrome P450 enzymes and animal microsomes, *Xenobiotica* 39 (2009) 11–21, <http://dx.doi.org/10.1080/00498250802512830>.
- [33] D.J. Dix, K.A. Houck, M.T. Martin, A.M. Richard, R.W. Setzer, R.J. Kavlock, The ToxCast program for prioritizing toxicity testing of environmental chemicals, *Toxicol. Sci.* 95 (2007) 5–12, <http://dx.doi.org/10.1093/toxicol/kfl103>.
- [34] ToxCast, iCSS ToxCast Dashboard, 2016. <https://actor.epa.gov/dashboard/> (accessed October 13, 2016).
- [35] K.L. Dionisio, A.M. Frame, M.-R. Goldsmith, J.F. Wambaugh, A. Liddell, T. Cathey, D. Smith, J. Vail, A.S. Emstoft, P. Fantke, O. Joliet, R.S. Judson, Exploring consumer exposure pathways and patterns of use for chemicals in the environment, *Toxicol. Reports* 2 (2015) 228–237, <http://dx.doi.org/10.1016/j.toxrep.2014.12.009>.
- [36] P.P. Egeghy, R. Judson, S. Gangwal, S. Mosher, D. Smith, J. Vail, E.A. Cohen, Hubal, The exposure data landscape for manufactured chemicals, *Sci. Total Environ.* 414 (2012) 159–166, <http://dx.doi.org/10.1016/j.scitotenv.2011.10.046>.
- [37] ECHA, Guidance on information requirements and chemical safety assessment, 2008. https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf (accessed December 8, 2016).
- [38] OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, 2, 2007, pp. 1–154. [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2007\)2&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en) (accessed December 8, 2016).
- [39] OpenTox, Final database with additional content, 2011. <http://opentox.org/data/documents/development/opentoxreports/opentoxreportd34/view> (accessed October 14, 2016).
- [40] Reach, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006, Off. J. Eur. Communities, L 269, 2006, pp. 1–15. <http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02006R1907-20140410&from=EN> (accessed October 14, 2016).
- [41] C.L. Mellor, F.P. Steinmetz, M.T.D. Cronin, Using molecular initiating events to develop a structural alert based screening workflow for nuclear receptor ligands associated with hepatic steatosis, *Chem. Res. Toxicol.* 29 (2016) 203–212, <http://dx.doi.org/10.1021/acs.chemrestox.5b00480>.
- [42] R. Huang, N. Southall, Y. Wang, A. Yaggar, P. Shinn, A. Jadhav, D.-T. Nguyen, C.P. Austin, The NCGC pharmaceutical collection: a comprehensive resource of

- clinically approved drugs enabling repurposing and chemical genomics, *Sci. Transl. Med.* 3 (2011) 80ps16, <http://dx.doi.org/10.1126/scitranslmed.3001862>.
- [43] S.J. Shukla, D.-T. Nguyen, R. MacArthur, A. Simeonov, W.J. Frazee, T.M. Hallis, B. D. Marks, U. Singh, H.C. Eliason, J. Printen, C.P. Austin, J. Inglese, D.S. Auld, Identification of pregnane X receptor ligands using time-resolved fluorescence resonance energy transfer and quantitative high-throughput screening, *Assay Drug Dev. Technol.* 7 (2009) 143–169, <http://dx.doi.org/10.1089/adt.2009.193>.
- [44] QSAR User Manual for the Danish (Q)SAR Database, 2015. http://qsar.db.food.dtu.dk/Danish_QSAR_Database_Draft_User_manual.pdf (accessed October 14, 2016).
- [45] QSARDB, Danish (Q)SAR Database, 2015. <http://qsar.food.dtu.dk/> (accessed October 14, 2016).
- [46] Leadscape, Leadscape, Inc., 2016. <http://www.leadscope.com/> (accessed October 14, 2016).
- [47] G. Roberts, G.J. Myatt, W.P. Johnson, K.P. Cross, P.E. Blower, LeadScope \ddagger : software for exploring large sets of screening data, *J. Chem. Inf. Comput. Sci.* 40 (2000) 1302–1314, <http://dx.doi.org/10.1021/ci0000631>.
- [48] L.G. Valerio, C. Yang, K.B. Arvidson, N.L. Kruhlik, A structural feature-based computational approach for toxicology predictions, *Expert Opin. Drug Metab. Toxicol.* 6 (2010) 505–518, <http://dx.doi.org/10.1517/17425250903499286>.
- [49] J.A. Cooper II, R. Saracci, P. Cole, Describing the validity of carcinogen screening tests, *Br. J. Cancer* 39 (1979) 87–89.
- [50] X. Ding, J.L. Staudinger, Repression of PXR-mediated induction of hepatic CYP3A gene expression by protein kinase C, *Biochem. Pharmacol.* 69 (2005) 867–873, <http://dx.doi.org/10.1016/j.bcp.2004.11.025>.
- [51] W. Lin, J. Wu, H. Dong, D. Bouck, F. Zeng, T. Chen, Cyclin-dependent kinase 2 negatively regulates human pregnane X receptor-mediated CYP3A4 gene expression in HepG2 liver carcinoma Cells*, *J. Biol. Chem.* 283 (2008) 30650–30657, <http://dx.doi.org/10.1074/jbc.M806132200>.
- [52] M.D.M. AbdulHameed, D.L. Ippolito, A. Wallqvist, Predicting rat and human pregnane X receptor activators using bayesian classification models, *Chem. Res. Toxicol.* (2016), <http://dx.doi.org/10.1021/acs.chemrestox.6b00227>.
- [53] F.P. Steinmetz, S.J. Enoch, J.C. Madden, M.D. Nelms, N. Rodriguez-sanchez, P.H. Rowe, Y. Wen, M.T.D. Cronin, Methods for assigning confidence to toxicity data with multiple values – Identifying experimental outliers, *Sci. Total Environ.* 482–483 (2014) 358–365, <http://dx.doi.org/10.1016/j.scitotenv.2014.02.115>.
- [54] M. Gütlein, C. Helma, A. Karwath, S. Kramer, A large-scale empirical evaluation of cross-validation and external test set validation in (Q)SAR, *Mol. Inform.* 32 (2013) 516–528, <http://dx.doi.org/10.1002/minf.201200134>.
- [55] B.L. Ingle, B.C. Veber, J.W. Nichols, R. Tornero-Velez, Informing the human plasma protein binding of environmental chemicals by machine learning in the pharmaceutical space: applicability domain and limits of predictability, *J. Chem. Inf. Model.* (2016), <http://dx.doi.org/10.1021/acs.jcim.6b00291>.
- [56] Y. Yin, D.T. Chang, C.M. Grulke, Y.-M. Tan, M.-R. Goldsmith, R. Tornero-Velez, Essential set of molecular descriptors for ADME prediction in drug and environmental chemical space, *Research*, (2014), <http://dx.doi.org/10.13070/rs.en.1.996>.
- [57] T.M. Martin, P. Harten, D.M. Young, E.N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, Does rational selection of training and test sets improve the outcome of QSAR modeling?, *J. Chem. Inf. Model.* 52 (2012) 2570–2578, <http://dx.doi.org/10.1021/ci300338w>.
- [58] F. Shah, N. Greene, Analysis of Pfizer compounds in EPA's ToxCast chemicals-assay space, *Chem. Res. Toxicol.* 27 (2014) 86–98, <http://dx.doi.org/10.1021/tx400343t>.
- [59] C.-H. Ngan, D. Beglov, A.N. Rudnitskaya, D. Kozakov, D.J. Waxman, S. Vajda, The structural basis of pregnane X receptor binding promiscuity, *Biochemistry* 48 (2009) 11572–11581, <http://dx.doi.org/10.1021/bi901578n>.
- [60] G. Luo, M. Cunningham, S. Kim, T. Burn, J. Lin, M. Sinz, G. Hamilton, C. Rizzo, S. Jolley, D. Gilbert, A. Downey, D. Mudra, R. Graham, K. Carroll, J. Xie, A. Madan, A. Parkinson, D. Christ, B. Selling, E. LeCluyse, L.-S. Gan, CYP3A4 induction by drugs: correlation between a pregnane X receptor reporter gene assay and CYP3A4 expression in human hepatocytes, *Drug Metab. Dispos.* 30 (2002) 795–804, <http://dx.doi.org/10.1124/dmd.30.7.795>.
- [61] T.A. Kocarek, E. Schuetz, P. Guzelian, Regulation of phenobarbital-inducible cytochrome P450 2B1/2 mRNA by lovastatin and oxysterols in primary cultures of adult rat hepatocytes, *Toxicol. Appl. Pharmacology* 120 (1993) 298–307, <http://dx.doi.org/10.1006/taap.1993.1115>.
- [62] C.A. LaLone, D.L. Villeneuve, D. Lyons, H.W. Helgen, S.L. Robinson, J.A. Swintek, T.W. Saari, G.T. Ankley, Sequence alignment to predict across species susceptibility (SeqAPASS): a web-based tool for addressing the challenges of cross-species extrapolation of chemical toxicity, *Toxicol. Sci.* 153 (2016) 228–245, <http://dx.doi.org/10.1093/toxsci/kfw119>.

3.3 QSAR Models for AhR Activation *In Vitro*

3.3.1 Study Report

A pilot study to explore the effect of rational selection of training set inactives on model predictive performance and coverage using a large imbalanced AhR activation dataset

Rosenberg, S.A.^a, Dybdahl, M.^{a1}, Wedebye, E.B.^{a1}, and Nikolov, N.G.^{a1}

a. Division of Diet, Disease Prevention and Toxicology, National Food Institute, Technical University of Denmark, Kemitorvet, Building 202, 2800 Kgs. Lyngby, Denmark

¹*Contributed equally*

Abbreviations: AD, applicability domain; AhR, aryl hydrocarbon receptor; AOP, Adverse Outcome Pathway; CYP, cytochrome P450; ER, estrogen receptor; FN, false negative; FP, false positive; HAH, halogenated aromatic hydrocarbons; HTS, high-throughput screenings; LPDM, Leadscape® Predictive Data Miner; MIE, molecular initiating event; PAH, polycyclic aromatic hydrocarbons; *q*HTS, *quantitative* HTS; QSAR, quantitative structure-activity relationship; SULT, sulfotransferase; TH, thyroid hormone; TN, true negative; TP, true positive; UGT, UDP-glucuronosyltransferase

1. Introduction

With the recent advances in *in vitro* assay technologies, data from high-throughput screenings (HTS) for molecular and cellular responses are becoming more and more common in public databases such as the PubChem database²¹ [1,2]. Such HTS datasets are often large, i.e. they can contain up to 100,000s of samples tested, and tend to be highly imbalanced towards many inactives [2,3]. Previously, data shortage has been one of the main limiting factors for developing robust global quantitative structure-activity relationship (QSAR) models. The availability of large but highly imbalanced HTS datasets for molecular and cellular responses to chemicals has introduced new challenges when building global QSARs [2,3]. The datasets with 100,000s of entries are generally too large for most QSAR software to handle in a computer- and time-efficient way, and the very imbalanced distribution of actives to inactives poses a problem for many training algorithms. One solution is therefore to select a subset to be used for QSAR training, e.g. with the aim of building models with good predictive performance and/or high coverage of future prediction sets. Suggestions on subset sampling and mining of large imbalanced HTS datasets have been published previously [2–4]. The predictive performance of a QSAR, i.e. how good it is at making correct and reliable predictions, is strongly influenced by the quality of the underlying experimental data and structures on which it has been trained [5,6]. For global QSARs, the size and balance of the training set, the distribution of training set structures in the chemical space as well as the definition of an applicability domain (AD) also play a role in a model's estimated predictive performance. Model coverage, also defined as the AD size, is the proportion of a prediction set for which the QSAR model can make predictions within the reliability established in the QSAR validation. Addition of structures to a training set can enhance the model's coverage and predictive performance, and the degree of coverage and predictive performance improvement will most likely depend of the number of structures added as well as their effect on the chemical space covered by the training set.

The aryl hydrocarbon receptor (AhR) is a ligand-dependent transcription factor that regulates the expression of genes, whose products are involved in multiple biological processes such as metabolism of endogenous and exogenous small molecules as well as regulation of organ development and the immune system [7]. Due to its wide and important biological involvement, AhR

²¹ <https://pubchem.ncbi.nlm.nih.gov/>

continues to be a popular research area²². Some of the best-characterized exogenous AhR ligands include dioxins, halogenated aromatic hydrocarbons (HAHs) and nonhalogenated polycyclic aromatic hydrocarbons (PAHs). Further studies have identified a structurally-diverse group of chemicals as AhR agonists [7]. Some of the genes regulated by AhR encode enzymes involved in phase I and II metabolism of exogenous as well as endogenous compounds. The two AhR-regulated cytochrome P450 (CYP) subtypes, CYP1A1 and CYP1B1, are among other things involved in phase I metabolism of estrogens [8–10]. AhR also regulates the expression of sulfotransferase (SULT) and UDP-glucuronosyltransferase (UGT) isoenzymes that are important in the catabolism of e.g. thyroid hormones (THs) and estrogens [11–13]. Thus exposure to man-made chemicals that interact with AhR can through upregulation of enzymes such as CYPs, UGTs and SULTs result in altered turnover of endogenous hormones and hereby potentially interfere with normal physiology and lead to adverse health effects. One example is given in an Adverse Outcome Pathway (AOP) (under development) that describes how the molecular initiating event (MIE) of chemical interaction with AhR upregulates TH catabolism and leads to reduced TH levels and can result in adverse neurodevelopmental outcomes [14]. AhR can also modulate the responsiveness of various hormone receptors [7]. Best understood is the cross-talk with the estrogen receptors (ERs), whose activity can be repressed by ligand-activated AhRs through sequestering of common co-activators/factors [7]. Similar types of cross-talk between AhR and other nuclear receptors and transcription factors are likely [7].

Due to the involvement of AhR in toxic responses to chemicals such as reduced TH levels and neurodevelopmental adverse outcomes [14], it is of high relevance to be able to identify and characterize chemical structures that activate AhR. A number of HTS *in vitro* assays for AhR interaction have been developed and applied to screen thousands of small molecules [15,16]. Such data have previously been used in the development of QSAR models for AhR activation, e.g. QSAR models developed from Tox21 HTS data under the Tox21 challenge in 2014 [15]. In the present study, a large PubChem dataset with 324,858 chemical structures probing the classical AhR-gene activation mechanism in a *quantitative* HTS (*qHTS*) *in vitro* assay was curated and used to prepare training and test sets to build and validate four global QSAR models. Corresponding data on luciferase interference, a potential artefact in the applied AhR activation assay, was taken into account to remove potentially false positive experimental results from the AhR activation dataset at the data curation step. Due to the high ratio of 204,513 AhR activation inactives to 925 actives in the curated dataset, we used this dataset to explore how a stepwise rational selection of inactives to expand training set size would affect the coverage and predictive performances of the QSAR models.

²² <http://www.sciencedirect.com/science/journal/aip/24682020>

2. Material and Methods

2.1 Experimental datasets

A dataset consisting of structure information and *q*HTS *in vitro* data for human AhR activation and luciferase interference was used when constructing training and test sets. All data were downloaded from PubChem. In total, 324,858 chemicals had been tested in a primary singlicate screening for AhR activation, i.e. AID 2796, and given a PubChem activity score of 0-100 as described elsewhere [16]. Of the 7,990 substances originally tested active in AID 2796, 2,281 had been retested in triplicate for AhR activation, i.e. AID 2845 [17], and of these, 1,982 were confirmed AhR activators, i.e. PubChem activity score of 10-100 [17]. The AhR activation *q*HTS *in vitro* assay applied in AID 2796 and AID 2845 is a luminescence-based assay using HepG2 cells stably transfected with AhR-dependent pGudLuc6.1-DRE plasmids [17]. Substances that activate AhR result in expression of the luciferase reporter gene, and the level of luciferase activity is an indirect measure of AhR activation [17]. Some substances can stabilize luciferase and increase its half-life resulting in its accumulation and a measured increase in luminescence signal [18], and such substances may be incorrectly interpreted as AhR activators in the applied AhR activation *q*HTS assay. We used experimental PubChem data from the luciferase inhibition/activation *q*HTS assay AID 5888342 [19] as a counterscreen to identify any such substances among the 1,982 confirmed AhR activators from AID 2845. We classified substances in AID 2845 with a PubChem activity score from 10 to 100 and a PubChem activity score of 0 in AID 588342 as active for AhR activation. Substances with a PubChem score of 0 in AID 2796 were classified as inactive for AhR activation. The remaining substances were classified as inconclusive for AhR activation.

2.2 Structure preparation and dataset splitting

The QSAR software applied in this study, Leadscape® Predictive Data Miner (LPDM), a component of Leadscape® Enterprise Server version 3.2.4, can handle organic chemical substances with a known and unambiguous 2D structure [20]. Briefly, we prepared calculation structures by first breaking ionic bonds and neutralizing the structures. Then we removed substances containing two or more organic components and structures with less than two carbon atoms from the dataset. Also, structures containing atoms not on the following list were removed: H, Li, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and I. Finally, structures with charges in their calculation structures were removed from the dataset. Canonized SMILES were generated for the remaining calculation structures in the dataset so that they were described following the same algorithm (Figure 1, pink box) and these constituted the QSAR-ready structures that were used for further processing.

In the next step, identical QSAR-ready structures in the dataset were identified and their experimental results, as classified above, were compared. For identical structures with concordant activities, only one of the structures was kept in the dataset, while if a group of identical structures had discrepant activities then the whole group was removed from the dataset (Figure 1, pink box). After structure preparation and duplicates removal, the dataset was split as follows. Among the 925 AhR activation actives in the dataset, 10% were randomly selected to be used in a test set. This resulted in 93 test set actives and 832 training set actives. From the 204,513 QSAR-ready inactive in the dataset, we randomly selected 50,000 of the structures (to be called the '50K set' below) to be used in the model development steps as explained below, while the remaining 154,513 structures were included in the test set (Figure 1, pink box).

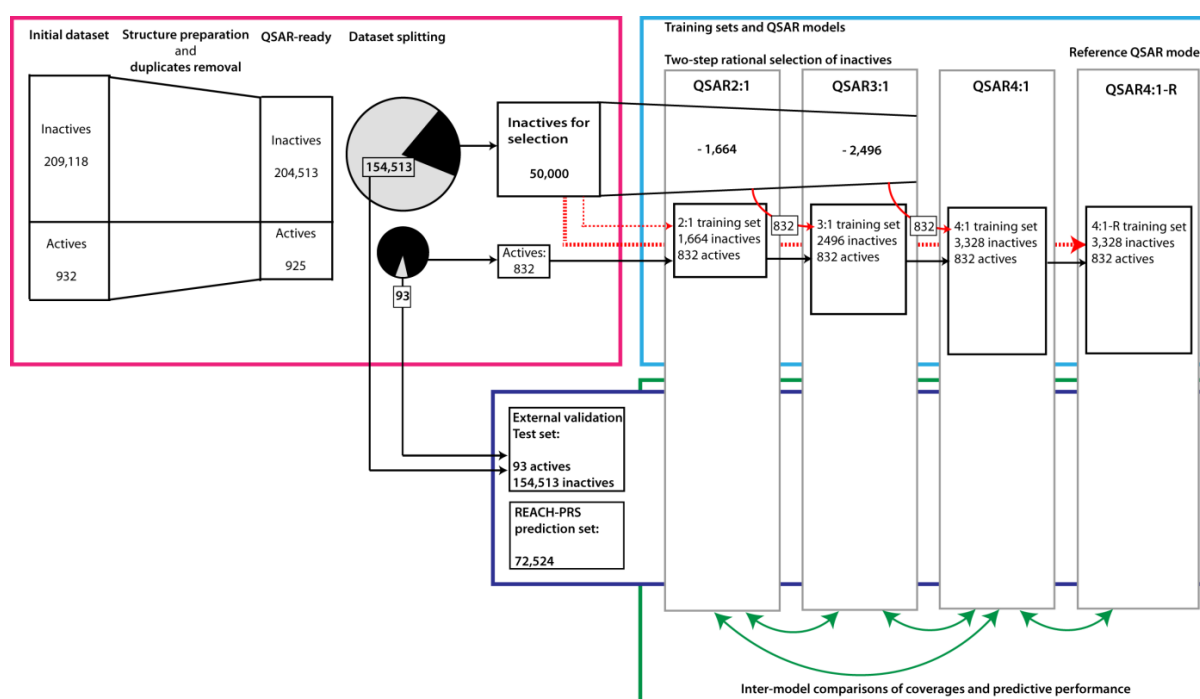


Figure 1. An overview of the workflow. Pink box: the steps of data curation and preparation of a test set and a dataset for training set construction. Light blue box: the steps of training set inactives selections and model building. Dark blue box: predicting the test set for external validation and the REACH-PRS set in the four models. Green box: inter-model comparisons of the predictive performances from the external validations and the coverages of the REACH-PRS set.

2.3 Applicability domain definition

The definition of the AD applied in this study consists of two components: 1) the definition of a structural domain in LPDM, and 2) a DTU Food in-house class probability refinement on the output from LPDM:

1) For a query compound to be within LPDM's structural domain it is required that: it has at least 30% Tanimoto similarity with a training set compound, all molecular descriptors used in the model can be calculated and it contains at least one structural feature used in the model [21]. The 30%

Tanimoto similarity was a default cut-off in the LPDM software. For a test compound outside this structural domain no prediction call, i.e. active/inactive, is generated by LPDM. For test compounds within the LPDM structural domain, a positive prediction probability, p , between 0 and 1, is given together with the prediction call; actives having a $p \geq 0.5$ and inactives having a $p < 0.5$ [21].

2) The DTU Food class probability refinement served to exclude the likely less reliable predictions, i.e. those with a positive prediction probability close to the cutoff $p = 0.5$. For predictions to be within the AD we required a $p \geq 0.7$ for active prediction calls (POS_IN) and a $p \leq 0.3$ for inactive prediction calls (NEG_IN). Predictions within the LPDM structural domain but with an associated positive prediction probability in the interval $0.3 < p < 0.5$ (NEG_OUT) and $0.5 \leq p < 0.7$ (POS_OUT) are defined as out of AD.

2.4 QSAR Modeling

In this study, we used the commercial software LPDM to build QSAR models. Briefly, upon dataset import LPDM calculates nine molecular descriptors (AlogP, Hydrogen Bond Acceptors and Donors, Lipinski Score, Molecular Weight, Parent Atom Number, Parent Molecular Weight, Polar Surface Area, Rotatable Bonds) and performs a systematic sub-structural analysis using a template library of more than 27,000 pre-defined structural keys for each chemical structure in the dataset [22]. For QSAR modeling in LPDM, the molecular descriptors and structural features are included in a default preliminary descriptor set. From the preliminary descriptor set, an automatic descriptor pre-selection procedure in LPDM selects the top 30% descriptors according to Yates X^2 -test for a binary response variable. For training sets with a binary response variable, a predictive model is built using the pre-selected descriptors in a partial logistic regression (PLR) with further selection of descriptors in an iterative procedure, and selection of the optimum number of PLR factors based on minimizing the predictive residual sum of squares. LPDM has the option of building composite models, a type of ensemble models, for training sets with an imbalanced distribution of actives and inactives [23]. With this option a number of sub-models are created by specifying the desired ratio of actives to inactives per sub-model training set. The positive prediction probability (see 2.3) for a query chemical from a composite model is defined as the average of the positive prediction probabilities from all sub-models having the test chemical in their structural domain [21].

To first find the maximal modeling capacity in LPDM of the present dataset, we did a series of modeling experiments using training sets with different ratios of the 832 actives and randomly selected inactives from the 50K set. The training set with a ratio of 4:1, i.e. consisting 3,328 inactives randomly selected and the 832 actives, was the largest imbalanced training set that LPDM could

efficiently model. This 4:1 training set was later used for building a reference model for evaluating the effect of the rational selection steps described below.

After determining the maximum training set inactive:active ratio we started to construct a 4:1 training set using a two-step rational selection procedure. We first created a training set with an inactive:active ratio of 2:1 that consisted of the 832 actives and 1,664 (i.e., twice the 832 actives) inactives selected randomly from the 50K set of inactives (Figure 1, light blue box). The 2:1 training set was modeled in LPDM using three QSAR modeling approaches, which all underwent a 10 times 20%-out LPDM cross-validation:

- 1) A single model, i.e. a non-composite model using the full training set
- 2) A composite model, with sub-models from balanced sub training sets and equal weight
- 3) A composite 'cocktail' model, combining the single model from 1) with the sub-models of the composite model from 2)

Since the main purpose in this study was to compare the predictive performances and coverages between models built from training sets constructed using two different selection approaches, we decided that all models should be built using the same modeling approach. Based on the LPDM cross-validation results the best performing modeling approach was selected and the selected model was closed and named QSAR2:1. Then the 50K set minus the inactive structures in the 2:1 training set, i.e. 48,336 inactive structures, were predicted in QSAR2:1 (Figure 1, light blue box). From these predictions, 832 new inactives were selected and added to the 2:1 training to constitute a 3:1 training set as follows. The rational selection was done by selecting one fourth, corresponding to 208 structures, randomly from each of the four prediction outcome areas (defined in 2.3):

1. out of LPDM structural domain
2. POS_OUT
3. NEG_OUT
4. POS_IN, i.e. here false positive (FP) predictions

The addition of structures from 1. through 3. mainly served to increase chemical space of the subsequent training set with the purpose of increasing the AD and model coverage. The structures with POS_IN predictions, i.e. 4., were added with the purpose to improve the ability of the model algorithm to avoid deriving false activity features and thereby reduce its tendency to make FP predictions. A similar but smaller effect on performance was expected from addition of the POS_OUT (2.) and NEG_OUT (3.) selected structures.

The 3:1 training set was used for building a QSAR model using the selected modeling approach, and the model was closed and named QSAR3:1. The 50K minus the 3:1 training set inactive structures, i.e. 47,504 inactive structures, were then predicted in QSAR3:1 and from the predictions, 832 inactives were selected as described above and added to the 3:1 training set to constitute a 4:1 training set (Figure 1, light blue box). Again, the 4:1 training set was used for building a QSAR model using the selected modeling approach and the model was closed and named QSAR4:1. To have a reference model to evaluate the effect of the rational selection steps against, the 4:1 training set with the inactives randomly selected from the 50K set were used for building a model using the selected modeling approach. This model was closed and named QSAR4:1-R.

2.5 Validation of the QSAR models

All four selected and closed models, QSAR2:1, QSAR3:1, QSAR4:1 and QSAR4:1-R, had during their development undergone a 10 times 20%-out cross-validation procedure in LPDM. The LPDM cross-validation applies the LPDM structural domain only and is not a true cross-validation as the algorithm transfers knowledge from the full training set model to the smaller cross-validation models. Therefore, the LPDM cross-validation results were only used in a relative manner to guide the selection of the modeling approach (see 2.4) and not to estimate absolute predictive performance. To assess the models predictive performances, the four closed models were subjected to an external validation using the test set of 93 AhR actives and 154,513 inactives (Figure 1, dark blue box). Sensitivity, specificity and balanced accuracy were calculated for the test set predictions within the defined AD. Sensitivity is the percentage of experimental actives correctly predicted, specificity is the percentage of the experimental inactives correctly predicted, and balanced accuracy is the average of the sensitivity and specificity. The coverage of the test set, i.e. the percentage of how many of the predicted test set structures that had predictions within the defined AD, was also calculated for all four QSAR models.

2.6 Screening of 72,524 REACH substances for AhR activation

An EU collection of 72,524 substances from the REACH pre-registered substances (PRS) list extracted from the online Danish (Q)SAR Database structure set [24,25] was screened through the four AhR activation QSAR models (Figure 1, dark blue box). The 72,524 QSAR-ready structures were originally curated from deliverable 3.4 of the OpenTox EU project [26] and had previously been processed through the structure preparation steps described in 2.2. The proportion of the 72,524 QSAR-ready REACH-PRS structures predicted within the defined AD of each of the four QSAR models, respectively, as well as the activity distributions of the predictions were calculated.

2.7 Comparison of model coverages and predictive performances

To uncover the effect of the two-step rational selection of inactives for the QSAR4:1 training set, an analysis of the coverages of the REACH-PRS set and the test set in the four models was performed. The results from the external validation of the four models using the test set were also compared to assess the effect of the stepwise rational selection procedure with regard to predictive performance. The analyses and comparisons were focusing on QSAR4:1 versus QSAR4:1-R as well as between the intermediate models QSAR2:1 and QSAR3:1 versus QSAR4:1 (Figure 1, green box).

3. Results and Discussion

Here we describe a pilot study to explore how a large and highly inactive-imbalanced dataset could be used for developing global QSAR models with optimized coverages and predictive performances.

3.1 The datasets

According to our classification of AhR actives and inactives described in 2.1 the initial dataset contained 932 actives and 209,118 inactives. During the structure preparation and duplicates handling in 2.2, a total of 4,612 structures were removed from the dataset, 2,909 due to the structural QSAR criteria and 1,703 due to structural duplicates, none of which due to conflicting experimental results (Figure 1, pink box). The number of QSAR-ready structures and the distribution of active and inactive experimental results in the full curated dataset, the test set, the 50K set for training set selection of inactives as well as the four training sets are summarized in Table 1.

Table 1. Overview of the datasets and their distributions of active and inactive experimental results.

Dataset overview	Actives	Inactives	Total
Full dataset	925	204,513	205,438
Test set	93	154,513	154,606
50K set	0	50,000	50,000
2:1 training set	832	1,664	2,496
3:1 training set	832	2,496	3,328
4:1 training set	832	3,328	4,160
4:1-R training set	832	3,328	4,160

3.2 Selection of model building approach

The 2:1 training set was used for building three QSAR models applying three different modeling approaches in LPDM. Their LPDM cross-validation results are given in Table 2. These results were used for selecting the modeling approach and not for estimating model predictive performance. As can be seen from Table 2, all three modeling approaches showed similar balanced accuracies from 81.3% to 83.7% in the 10 times 20%-out LPDM cross-validation. The lower LPDM sensitivity of the single model was expected due to the imbalance of the training set. The 2:1 training set composite 'cocktail' model 3) was the modeling approach that produced the highest number of both true

positive (TP) and true negative (TN) predictions and it resulted in more moderate numbers of FP and false negative (FN) predictions compared to the two other approaches. Based on these numbers, and on the fact that the composite modeling approach in LPDM is designed to handle imbalanced training sets, we selected the composite modeling approach 3) for future modeling of the remaining training sets, 3:1, 4:1 and 4:1-R.

Table 2. The results from the 10 times 20%-out LPDM cross-validations of the three modeling approaches applied on the 2:1 training set.

2:1 training set	Predictions in LPDM structural domain				Statistical parameters		
Modeling approach	TP	TN	FP	FN	Sensitivity, %	Specificity, %	Balanced accuracy, %
1) Single	587	1423	154	224	72.4	90.2	81.3
2) Composite	666	1266	262	122	84.5	82.9	83.7
3) 'Cocktail'	670	1427	224	162	80.5	86.4	83.5

TP = true positive, TN = true negative, FP = false positive, FN = false negative

3.3 Predictive performance assessment by external validation

After building the four models as described in 2.4, they were all subjected to external validation with the test set. In Table 3, the external validation results from the four QSAR models are given. An overall increase was seen when comparing the predictive performances from the external validations of QSAR2:1, QSAR3:1 and QSAR4:1. The stepwise rational selection with addition of inactives to the 2:1 and 3:1 training sets gave a total increase in specificity of 7%, i.e. from 90.2% in QSAR2:1 to 97.2% in QSAR4:1. The sensitivity was more or less unaffected and ranged from 83.6% to 85.7% without a trend between the models, and these small differences in the sensitivities are likely mainly due to noise.

Table 3. The results from the external validation of the four models including model coverage of the test set.

External validation		QSAR2:1	QSAR3:1	QSAR4:1	QSAR4:1-R
Statistical parameters, %	Sensitivity (TP/(TP+FN))	85.7	83.6	85.1	89.8
	Specificity (TN/(TN+FP))	90.2	95.3	97.2	91.6
	Balanced accuracy	88.0	89.5	91.2	90.7
POS_IN	TP	60	46	40	53
	FP	11,605	5,652	3,320	10,017
NEG_IN	TN	107,377	114,165	115,045	109,320
	FN	10	9	7	6
Coverage	Of 93 actives (%)	70 (75.3)	55 (59.1)	47 (50.5)	59 (63.4)
	Of 154,513 inactives (%)	118,982 (77.0)	119,817 (77.5)	118,365 (76.6)	119,337 (77.2)
	In total	119,052	199,872	118,412	119,396
	(%)	(77.0)	(77.5)	(76.6)	(77.2)

The test set consisted of 93 actives and 154,513 inactives. TP = true positive, TN = true negative, FP = false positive, FN = false negative

A comparison of the external validation statistical parameters from QSAR4:1 and QSAR4:1-R showed that the QSAR4:1 model had a higher specificity, i.e. 97.2% versus 91.6%, but a lower sensitivity, i.e. 85.1% versus 89.8%, than the QSAR4:1-R (Table 3). The positive effect on the specificity was an

expected result from the procedure of rational addition of inactives selected among the POS_IN and POS_OUT predictions produced by the preceding models. Inclusion of these structures with false positive predictions in the training set can help the subsequent model train on a more representative chemical space and thereby make more correct predictions.

3.3 Model coverages

Another focus of this study was to explore how the selection of inactives for the training sets would affect future model coverages. In Table 3, the coverages of the test set in the four models are given and as can be seen all models showed test set coverages of 76.6% to 77.5%. Thus, no effect on overall test set coverage was seen from the two-step rational versus the random selection. Although the inter-model total coverages are similar, there are clear differences in the absolute number of TP, TN, FP and FNs, respectively, produced from the four models (see Table 3). The QSAR4:1 and QSAR4:1-R coverages of the small number of 93 test set actives were 50.5% (47/93) and 63.4% (59/93), respectively. Due to the low active-to-inactive ratio in the test set, i.e. 93 actives to 154,513 inactives, the differences in the coverages of actives between the models are blurred in the total coverage measures (Table 3).

Besides screening the test set structures in the four models, the REACH-RS inventory of 72,524 man-made chemicals was also predicted by the models. The prediction and coverage results from the REACH-PRS screening can be found in Table 4. In Figure 2, the coverages of the REACH-PRS are shown.

Table 4. Overview of the screening results from the REACH-PRS set

REACH-PRS screening	QSAR2:1	QSAR3:1	QSAR4:1	QSAR4:1-R
Coverage (%)	31,611 (43.6)	40,418 (55.7)	46,261 (63.8)	39,698 (54.7)
POS_IN	2,744	1,483	1,269	2,148
NEG_IN	28,867	38,935	44,992	37,550

When comparing the coverages of the REACH-PRS set in the two intermediate models QSAR2:1 and QSAR3:1 to QSAR4:1, a total increase in coverage of 20% can be observed (Figure 2 and Table 4). This increase was an expected effect of the gradual increase in training set size, and was especially an effect of the large increase in NEG_IN predictions relative to the fall in POS_IN predictions (Table 4). Despite the same number of actives and inactives in the QSAR4:1 and QSAR4:1-R training sets, the coverage of REACH-PRS was almost 10% larger in QSAR4:1, which is most likely an effect of the rational selection steps. Also here, QSAR4:1 produced more NEG_IN predictions, i.e. 44,992 versus 37,550, with a smaller absolute decrease in its number of POS_IN outputs, i.e. 1,269 versus 2,148, relative to QSAR4:1-R.

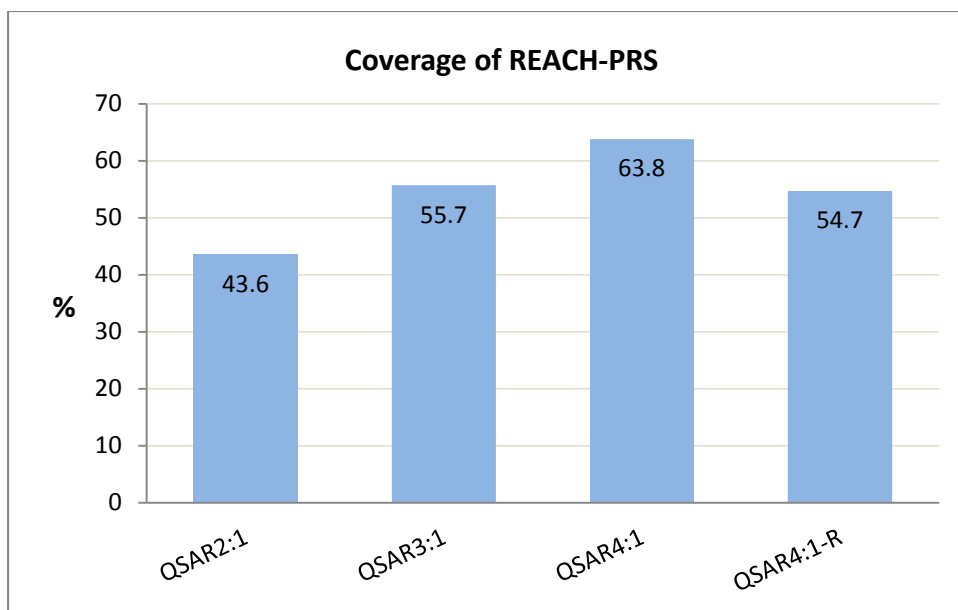


Figure 2. Coverage of the REACH-PRS set in the four QSAR models.

The more NEG_IN predictions produced by QSAR4:1 are likely a result of an increased structural diversity of inactives in the rational selected training set. This increase in structural diversity and the AD is mainly driven by the addition of structures with predictions out of LPDM structural domain (1.) in the preceding model as well as adding structures with NEG_OUT predictions that may have helped the subsequent model make more clear predictions, i.e. NEG_IN, for these types of structures. The addition of 50K inactive structures with false POS_IN and POS_OUT predictions in the intermediate models has likely helped the QSAR4:1 model reduce its rate of FP predictions, and is part of the reason for the smaller number of POS_IN REACH-PRS predictions generated from QSAR4:1. However, since the rational addition of structures was only aimed at increasing the number and diversity of inactive structures in the training set without a corresponding increase in training set actives, the addition of structures in the POS_IN and POS_OUT prediction areas has also resulted in a sacrifice of the number of TP predictions produced by QSAR4:1. This can also be seen in the results from the test set, where QSAR4:1 resulted in 40 TP predictions out of the 93 test set actives as opposed to the 53 TP predictions from QSAR4:1-R (Table 3).

Overall, these results indicate that the rational selection procedure of training set inactives for QSAR4:1 has produced a model with enlarged coverage of the large REACH-PRS prediction set (from 54.7% to 63.8%). The same effect was for unknown reasons not seen for the test set, instead a reduction in the coverage of the 93 actives (from 63% to 51%) was observed. The QSAR4:1 model according to the external validations produced the highest number of TNs but also the fewest TPs. Depending on the purpose of the QSAR screening, the four models may serve different aims. If the

QSAR screening is for example aiming at finding as many TP's as possible at the expense of a higher number of FP's, then the external validation indicates that QSAR2:1 is the best model.

4. Conclusions

Overall, the external validations showed that all four models had high predictive performances with balanced accuracies of 88.0% to 91.2%. From this pilot study, we can conclude that the stepwise rational selection of training set inactive structures from a very large and imbalanced datasets improved model specificity, i.e. ability to correctly predict the inactives, from 91.6% to 97.2% compared to random selection. The coverage improvement effect of the rational selection depended on the constitution of the prediction set, and here we saw an approximately 10% coverage increase of the REACH-PRS set but no improvement in test set coverage.

References

- [1] A. Tropsha, Best Practices for QSAR Model Development, Validation, and Exploitation, *Mol. Inform.* 29 (2010) 476–488. doi:10.1002/minf.201000061.
- [2] A. V. Zakharov, M.L. Peach, M. Sitzmann, M.C. Nicklaus, QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem, *J. Chem. Inf. Model.* 54 (2014) 705–712. doi:10.1021/ci400737s.
- [3] Q. Li, Y. Wang, S.H. Bryant, A novel method for mining highly imbalanced high-throughput screening data in PubChem, *Bioinformatics.* 25 (2009) 3310–3316. doi:10.1093/bioinformatics/btp589.
- [4] U. Norinder, S. Boyer, Binary classification of imbalanced datasets using conformal prediction, *J. Mol. Graph. Model.* 72 (2017) 256–265. doi:10.1016/j.jmgm.2017.01.008.
- [5] D. Fourches, E. Muratov, A. Tropsha, Curation of chemogenomics data, *Nat. Chem. Biol.* 11 (2015) 535–535. doi:10.1038/nchembio.1881.
- [6] F.P. Steinmetz, S.J. Enoch, J.C. Madden, M.D. Nelms, N. Rodriguez-Sanchez, P.H. Rowe, Y. Wen, M.T.D. Cronin, Methods for assigning confidence to toxicity data with multiple values — Identifying experimental outliers, *Sci. Total Environ.* 482–483 (2014) 358–365. doi:10.1016/j.scitotenv.2014.02.115.
- [7] M.S. Denison, A.A. Soshilov, G. He, D.E. DeGroot, B. Zhao, Exactly the Same but Different: Promiscuity and Diversity in the Molecular Mechanisms of Action of the Aryl Hydrocarbon (Dioxin) Receptor, *Toxicol. Sci.* 124 (2011) 1–22. doi:10.1093/toxsci/kfr218.
- [8] A.F. Badawi, E.L. Cavalieri, E.G. Rogan, Role of human cytochrome P450 1A1, 1A2, 1B1, and 3A4 in the 2-, 4-, and 16 α -hydroxylation of 17 β -estradiol, *Metabolism.* 50 (2001) 1001–1003. doi:10.1053/meta.2001.25592.
- [9] C.P. Martucci, J. Fishman, P450 enzymes of estrogen metabolism, *Pharmacol. Ther.* 57 (1993) 237–257. doi:10.1016/0163-7258(93)90057-K.
- [10] Y. Tsuchiya, M. Nakajima, T. Yokoi, Cytochrome P450-mediated metabolism of estrogens and its regulation in human, *Cancer Lett.* 227 (2005) 115–124. doi:10.1016/j.canlet.2004.10.007.
- [11] K.M. Crofton, Thyroid disrupting chemicals: mechanisms and mixtures, *Int. J. Androl.* 31 (2008) 209–223. doi:10.1111/j.1365-2605.2007.00857.x.

- [12] C. Guillemette, A. Bélanger, J. Lépine, Metabolic inactivation of estrogens in breast tissue by UDP-glucuronosyltransferase enzymes: an overview, *Breast Cancer Res.* 6 (2004) 246–254. doi:10.1186/bcr936.
- [13] A.J. Murk, E. Rijntjes, B.J. Blaauboer, R. Clewell, K.M. Crofton, M.M.L. Dingemans, J. David Furlow, R. Kavlock, J. Köhrle, R. Opitz, T. Traas, T.J. Visser, M. Xia, A.C. Gutleb, Mechanism-based testing strategy using in vitro approaches for identification of thyroid hormone disrupting chemicals, *Toxicol. Vitro.* 27 (2013) 1320–1346. doi:10.1016/j.tiv.2013.02.012.
- [14] AOP-8, Upregulation of Thyroid Hormone Catabolism via Activation of Hepatic Nuclear Receptors, and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/8> (accessed March 13, 2017).
- [15] R. Huang, M. Xia, D.-T. Nguyen, T. Zhao, S. Sakamuru, J. Zhao, S.A. Shahane, A. Rossoshek, A. Simeonov, Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs, *Front. Environ. Sci.* 3 (2016) 1–9. doi:10.3389/fenvs.2015.00085.
- [16] National Center for Biotechnology Information, PubChem BioAssay Database; AID=2796, (n.d.). <https://pubchem.ncbi.nlm.nih.gov/bioassay/2796> (accessed March 5, 2017).
- [17] National Center for Biotechnology Information, PubChem BioAssay Database; AID=2845, (n.d.). <https://pubchem.ncbi.nlm.nih.gov/bioassay/2845> (accessed March 5, 2017).
- [18] J.F. Thompson, L.S. Hayes, D.B. Lloyd, Modulation of firefly luciferase stability and impact on studies of gene regulation, *Gene.* 103 (1991) 171–177. doi:10.1016/0378-1119(91)90270-L.
- [19] National Center for Biotechnology Information, PubChem BioAssay Database; AID=588342, (n.d.). <https://pubchem.ncbi.nlm.nih.gov/bioassay/588342> (accessed March 5, 2017).
- [20] Leadscope, Leadscope, Inc, (2016). <http://www.leadscope.com/> (accessed March 23, 2017).
- [21] L.G. Valerio, C. Yang, K.B. Arvidson, N.L. Kruhlak, A structural feature-based computational approach for toxicology predictions, *Expert Opin. Drug Metab. Toxicol.* 6 (2010) 505–518. doi:10.1517/17425250903499286.
- [22] G. Roberts, G.J. Myatt, W.P. Johnson, K.P. Cross, P.E. Blower, LeadScope † : Software for Exploring Large Sets of Screening Data, *J. Chem. Inf. Comput. Sci.* 40 (2000) 1302–1314. doi:10.1021/ci0000631.
- [23] L. Breiman, Bagging Predictors, *Mach. Learn.* 24 (1996) 123–140. doi:10.1023/A:1018054314350.
- [24] QSARDB, Danish (Q)SAR Database, (2015). <http://qsar.food.dtu.dk/> (accessed March 14, 2017).
- [25] S.A. Rosenberg, M. Xia, R. Huang, N.G. Nikolov, E.B. Wedebye, M. Dybdahl, QSAR development and profiling of 72,524 REACH substances for PXR activation and CYP3A4 induction, *Comput. Toxicol.* 1 (2017) 39–48. doi:10.1016/j.comtox.2017.01.001.
- [26] OpenTox, Final database with additional content, (2011). <http://opentox.org/data/documents/development/opentoxreports/opentoxreportd34/view> (accessed October 14, 2016).

3.4 The Collaborative Estrogen Receptor Activity Prediction Project

3.4.1 Introduction

The Collaborative Estrogen Receptor Activity Prediction Project, abbreviated CERAPP, was initiated in 2013 by the U.S. EPA NCCT under the Endocrine Disruptor Screening Program (EDSP) laid out in 1998 [1–3]. In EDSP, a two-tiered approach is applied to screen a universe of around 10,000 chemicals for their potential to be endocrine disruptors. The Tier 1 screening consists of a battery of 11 endocrine-related *in vitro* and *in vivo* assays [4] that would cost around 1,000,000 USD/chemical, use a minimum of 520 animals/chemical and have a throughput of approximately 50 chemicals/year [3,5]. This challenge initiated the idea of a pre-tier 1 filter [6]. The aim of CERAPP was to use structure-based computer models to predict the full EDSP universe for estrogen receptor (ER) activity to aid in prioritizing EDSP chemicals for further Tier 1 testing. Due to the ease and low cost of running such models, the chemical universe for ER activity prediction was expanded to cover most of the man-made chemicals with potential human exposure in the United States [3,7]. The U.S. EPA NCCT contacted relevant research groups, including the QSAR team at DTU Food, to request them for participation in CERAPP, which in January 2016 resulted in a scientific publication [7], describing the methods and main results from the project.

Briefly, the CERAPP project is focused on the ER signaling pathway activation, an important mechanism of another area of the endocrine system and not directly considered a mechanism of thyroid hormone disruption. However, some common links between the ER signaling pathway and the thyroid system do exist, for example are some of the enzymes regulated by e.g. AhR and PXR involved in the synthesis and/or metabolism of both estrogens and THs [8,9]. Furthermore, cross-talk between ER and e.g. AhR may indirectly affect ER signaling and/or TH catabolism [10,11]. Also, estrogens have an effect on TH economy and function [12] and vice versa [13]. Thus, the thyroid and estrogen systems do interact [14] and together affect e.g. brain development and regulation of behavior [15].

3.4.2 My Contributions to CERAPP

My contributions to the CERAPP project consisted of building a binary global QSAR model in LDPM using the U.S. EPA NCCT provided ToxCast training set of 80 actives and 1,342 inactives for ER agonism and documenting the developed QSAR model in the QMRF format (Appendix). The QSAR team at DTU Food then predicted the U.S. EPA NCCT provided prediction set in the ER agonist QSAR model as well as in two previously built QSAR models for human ER α binding [16]. The predictions inside the defined AD (see AD definition in the QMRF, Appendix) of the ER agonism QSAR model as well as the QMRF were sent to U.S. EPA NCCT, who evaluated the model based on the predicted

evaluation set as described in the paper. Besides the work made for CERAPP, the model underwent a robust cross-validation (Appendix) and was applied for screening the REACH-PRS inventory of 72,524 chemical structures pre-registered under REACH [17]. The result from the cross-validation revealed a highly predictive model with a specificity of 94.4% and a sensitivity of 80.6%. Of the screened REACH-PRS set, 53,433 (73.7%) structures had predictions within the defined AD, and of these 4,918 were predicted ER agonists.

3.4.3 Published paper

A Section 508-conformant HTML version of this article is available at <http://dx.doi.org/10.1289/ehp.1510267>.

Research

CERAPP: Collaborative Estrogen Receptor Activity Prediction Project

Kamel Mansouri,^{1,2} Ahmed Abdelaziz,³ Aleksandra Rybacka,⁴ Alessandra Roncaglioni,⁵ Alexander Tropsha,⁶ Alexandre Varnek,⁷ Alexey Zakharov,⁸ Andrew Worth,⁹ Ann M. Richard,¹ Christopher M. Grulke,¹ Daniela Trisciuzzi,¹⁰ Denis Fourches,⁶ Dragos Horvath,⁷ Emilio Benfenati,⁵ Eugene Muratov,⁶ Eva Bay Wedebye,¹¹ Francesca Grisoni,¹² Giuseppe F. Mangiatordi,¹⁰ Giuseppina M. Incisivo,⁵ Huixiao Hong,¹³ Hui W. Ng,¹³ Igor V. Tetko,^{3,14} Ilya Balabin,¹⁵ Jayaram Kancherla,¹ Jie Shen,¹⁶ Julien Burton,⁹ Marc Nicklaus,⁸ Matteo Cassotti,¹² Nikolai G. Nikolov,¹¹ Orazio Nicolotti,¹⁰ Patrik L. Andersson,⁴ Qingda Zang,¹⁷ Regina Politi,⁶ Richard D. Beger,¹⁸ Roberto Todeschini,¹² Rui Li Huang,¹⁹ Sherif Farag,⁶ Sine A. Rosenberg,¹¹ Svetoslav Slavov,¹⁷ Xin Hu,¹⁹ and Richard S. Judson¹

¹National Center for Computational Toxicology, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA; ²Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, USA; ³Institute of Structural Biology, Helmholtz Zentrum Muenchen-German Research Center for Environmental Health (GmbH), Neuherberg, Germany; ⁴Chemistry Department, Umeå University, Umeå, Sweden; ⁵Environmental Chemistry and Toxicology Laboratory, IRCCS (Istituto di Ricovero e Cura a Carattere Scientifico)-Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy; ⁶Laboratory for Molecular Modeling, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; ⁷Laboratoire de Chimoinformatique, University of Strasbourg, Strasbourg, France; ⁸National Cancer Institute, National Institutes of Health (NIH), Department of Health and Human Services (DHHS), Bethesda, Maryland, USA; ⁹Institute for Health and Consumer Protection (IHCP), Joint Research Centre of the European Commission in Ispra, Ispra, Italy; ¹⁰Department of Pharmacy-Drug Sciences, University of Bari, Bari, Italy; ¹¹Division of Toxicology and Risk Assessment, National Food Institute, Technical University of Denmark, Copenhagen, Denmark; ¹²Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milan, Italy; ¹³Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration (USDA), Jefferson, Arizona, USA; ¹⁴BigChem GmbH, Neuherberg, Germany; ¹⁵High Performance Computing, Lockheed Martin, Research Triangle Park, North Carolina, USA; ¹⁶Research Institute for Fragrance Materials, Inc., Woodcliff Lake, New Jersey, USA; ¹⁷Integrated Laboratory Systems, Inc., Research Triangle Park, North Carolina, USA; ¹⁸Division of Systems Biology, National Center for Toxicological Research, USDA, Jefferson, Arizona, USA; ¹⁹National Center for Advancing Translational Sciences, NIH, DHHS, Bethesda, Maryland, USA

BACKGROUND: Humans are exposed to thousands of man-made chemicals in the environment. Some chemicals mimic natural endocrine hormones and, thus, have the potential to be endocrine disruptors. Most of these chemicals have never been tested for their ability to interact with the estrogen receptor (ER). Risk assessors need tools to prioritize chemicals for evaluation in costly *in vivo* tests, for instance, within the U.S. EPA Endocrine Disruptor Screening Program.

OBJECTIVES: We describe a large-scale modeling project called CERAPP (Collaborative Estrogen Receptor Activity Prediction Project) and demonstrate the efficacy of using predictive computational models trained on high-throughput screening data to evaluate thousands of chemicals for ER-related activity and prioritize them for further testing.

METHODS: CERAPP combined multiple models developed in collaboration with 17 groups in the United States and Europe to predict ER activity of a common set of 32,464 chemical structures. Quantitative structure-activity relationship models and docking approaches were employed, mostly using a common training set of 1,677 chemical structures provided by the U.S. EPA, to build a total of 40 categorical and 8 continuous models for binding, agonist, and antagonist ER activity. All predictions were evaluated on a set of 7,522 chemicals curated from the literature. To overcome the limitations of single models, a consensus was built by weighting models on scores based on their evaluated accuracies.

RESULTS: Individual model scores ranged from 0.69 to 0.85, showing high prediction reliabilities. Out of the 32,464 chemicals, the consensus model predicted 4,001 chemicals (12.3%) as high priority actives and 6,742 potential actives (20.8%) to be considered for further testing.

CONCLUSION: This project demonstrated the possibility to screen large libraries of chemicals using a consensus of different *in silico* approaches. This concept will be applied in future projects related to other end points.

CITATION: Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, Zakharov A, Worth A, Richard AM, Grulke CM, Trisciuzzi D, Fourches D, Horvath D, Benfenati E, Muratov E, Wedebye EB, Grisoni F, Mangiatordi GF, Incisivo GM, Hong H, Ng HW, Tetko IV, Balabin I, Kancherla J, Shen J, Burton J, Nicklaus M, Cassotti M, Nikolov NG, Nicolotti O, Andersson PL, Zang Q, Politi R, Beger RD, Todeschini R, Huang R, Farag S, Rosenberg SA, Slavov S, Hu X, Judson RS. 2016. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ Health Perspect* 124:1023–1033; <http://dx.doi.org/10.1289/ehp.1510267>

Introduction

There are tens of thousands of natural and synthetic chemical substances to which humans and wildlife are exposed (Dionisio et al. 2015; Egeghy et al. 2012; Judson et al. 2009). A subset of these compounds may disrupt normal functioning of the endocrine system and cause health hazards to both humans and ecological species (Birnbaum and Fenton 2003; Diamanti-Kandarakis

et al. 2009; Mahoney and Padmanabhan 2010; UNEP and WHO 2013). Endocrine-disrupting chemicals (EDCs) can mimic or interfere with natural hormones and alter their mechanisms of action at the receptor level, as well as interfere with the synthesis, transport, and metabolism of endogenous hormones (Diamanti-Kandarakis et al. 2009). Exposure to EDCs can lead to adverse health effects involving developmental, neurological,

reproductive, metabolic, cardiovascular, and immune systems in humans and wildlife (Colborn et al. 1993; Davis et al. 1993; Diamanti-Kandarakis et al. 2009).

The estrogen receptor (ER) is one of the most extensively studied targets related to the effects of EDCs (Mueller and Korach 2001; Shanle and Xu 2011). This concern about estrogen-like activity of man-made chemicals is because of their potential for negatively affecting reproductive function (Hileman 1994; Kavlock et al. 1996). The emergence of concerns about EDCs has resulted in regulations requiring assessment of chemicals for estrogenic activity [Adler et al. 2011; U.S. Environmental Protection Agency (EPA) 1996; U.S. Food and Drug Administration (FDA) 1996]. There are numerous *in vitro* and *in vivo* protocols to identify potential endocrine pathway-mediated effects of chemicals, including interactions with hormone receptors (Jacobs et al. 2008; Rotroff et al.

Address correspondence to R.S. Judson, U.S. EPA, National Center for Computational Toxicology, 109 T.W. Alexander Dr., Research Triangle Park, NC 27711 USA. Telephone: (919) 541-3085. E-mail: judson.richard@epa.gov

Supplemental Material is available online (<http://dx.doi.org/10.1289/ehp.1510267>).

I.B. is employed by Lockheed Martin, Research Triangle Park, NC. J.S. is employed by Research Institute for Fragrance Materials, Inc., Woodcliff Lake, NJ. Q.Z. is employed by Integrated Laboratory Systems, Inc., Research Triangle Park, NC.

The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency or the U.S. Food and Drug Administration.

The authors declare they have no actual or potential competing financial interests.

Received: 27 May 2015; Revised: 5 October 2015; Accepted: 8 February 2016; Published: 23 February 2016.

Mansouri et al.

2013; Shanle and Xu 2011; Sung et al. 2012). However, experimental testing of chemicals is expensive and time-consuming and currently impractical for application to the vast number of synthetic chemicals in use. Consequently, toxicological data and especially estrogenic activity data are available only for a limited number of compounds (Cohen Hubal et al. 2010; Egeghy et al. 2012; Judson et al. 2009).

The use of *in silico* approaches, such as quantitative structure–activity relationships (QSARs), is an alternative to bridge the lack of knowledge about chemicals when little or no experimental data are available. These structure-based methods are particularly appealing for their ability to predict toxicologically relevant end points quickly and at low cost (Muster et al. 2008; Vedani and Smiesko 2009). QSARs have been promoted and their use recognized since the pioneering work of Hansch in the 1960s (Fujita et al. 1964; Hansch et al. 1962; Hansch and Deusch 1966). The conceptual basis of QSARs is that chemicals with similar structures are hypothesized to exhibit similar behavior in living organisms. Thus, it should be possible to predict biological activity of new chemicals based on published experimental data. Several guidance documents to develop these modeling techniques are available in the literature (Dearden et al. 2009; Worth et al. 2005).

Recently, *in vitro* high-throughput screening (HTS) assays have emerged and become a viable tool for large-scale chemical testing (Judson et al. 2011; Kavlock and Dix 2010; Wetmore et al. 2012). HTS generates substantial amounts of data that can be used as a knowledge base to correlate chemical structures to their biological activities. Thus, QSARs can identify key structural characteristics in active chemicals and can use them to virtually screen large chemical libraries. Although there is concern about the overall accuracy of a QSAR model to predict the true activity of a particular chemical, accuracy can be high enough to use the results for prioritizing chemicals that are worth subjecting to experimental testing.

With the increasing number of new substances submitted to the U.S. EPA and the European Chemicals Agency for registration (~ 1,500 chemicals every year), there is a need to prioritize chemicals to speed up the process and lower the overall costs of testing (U.S. EPA 2015). The Toxicology Testing in the 21st Century (Tox21) collaboration and the U.S. EPA's Toxicity Forecaster (ToxCast™) projects are screening thousands of chemicals in HTS *in vitro* assays for a broad range of targets (Dix et al. 2007; Judson et al. 2010; Martin et al. 2010). Relevant to this paper, these two projects have in common ~ 1,800 chemicals tested in a battery of 18 ER-related assays (Huang et al. 2014; Judson et al. 2015).

This paper describes the results of the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP), which was organized by the National Center for Computational Toxicology at the U.S. EPA. The aim of the project was to use ToxCast™/Tox21 ER HTS assay data to develop and optimize predictive computational models, and to use their predictions to prioritize a large chemical universe of 32,464 unique chemical structures for further testing. Seventeen research groups from the United States and Europe participated in this project. These groups submitted 40 categorical models and 8 continuous models using different QSAR and structure-based approaches. Most of the newly developed models used a training set consisting of 1,677 chemicals, each assigned a potency score quantifying their ER agonist, antagonist, and binding activities, obtained from a computational network model that integrates data from 18 diverse ER HTS assays (Judson et al. 2015). All models were evaluated and weighted based on their prediction accuracy scores (including sensitivity and specificity) using ToxCast™/Tox21 HTS data, as well as an evaluation data set collected from different literature sources. To overcome the limitations of single models, all predictions were combined into a *consensus* model that classified the chemicals into active/inactive binders, agonists, and antagonists and provided estimates of their potency level relative to known reference chemicals.

Materials and Methods

Participants and Project Planning

The 17 international research groups that participated in this project are listed in alphabetic order in Table S1. The goals of the project, outlined in Table S2, were achieved in multiple steps, including chemical structure curation, experimental data preparation from the literature, modeling and prediction, model evaluation, consensus strategy development, and consensus modeling. Each step was assigned to a subgroup of participants according to their interests and areas of expertise.

Data Sets

Provided training set. The data that were suggested to be used by the participants as a training set to develop and optimize their models was derived from ToxCast™ and Tox21 programs (Dix et al. 2007; Huang et al. 2014; Judson et al. 2010). Concentration-response data from a collection of 18 *in vitro* HTS assays exploring multiple sites in the mammalian ER pathway were generated for 1,812 chemicals (Judson et al. 2015; U.S. EPA 2014c). This chemical library included 45 reference ER agonists and

antagonists (including negatives), as well as a wide array of commercial chemicals with known estrogen-like activity (Judson et al. 2015). A mathematical model was developed to integrate the *in vitro* data and calculate an area under the curve (AUC) score, ranging from 0 to 1, which is roughly proportional to the consensus AC50 value across the active assays (Judson et al. 2015). A given chemical was considered active if its agonist or antagonist score was higher than 0.01. In order to reduce the number of potential false positives this threshold can be increased to 0.1.

Prediction set. We identified > 50,000 chemicals [at the level of Chemical Abstracts Service Registry Number (CASRN)] for use in this project as a virtual screening library to be prioritized for further testing and regulatory purposes. This set was intended to include a large fraction of all man-made chemicals to which humans may be exposed. These chemicals were collected from different sources with significant overlap and cover a variety of classes, including consumer products, food additives, and human and veterinary drugs. The following list includes the sources used in this project:

- Chemicals with documented use, and therefore, with exposure potential (~ 43,000). Available in the U.S. EPA chemical product categories database (CPCat), which is part of the Aggregated Computational Toxicology Resource (ACToR) system (Dionisio et al. 2015; Judson et al. 2008, 2012; U.S. EPA 2014a).
- The Distributed Structure-Searchable Toxicity (DSSTox) (U.S. EPA 2014b). A list of ~ 15,000 curated chemical structures from multiple inventories of environmental interest. In particular, structures for all of the ToxCast™ and Tox21 chemicals are included.
- The Canadian Domestic Substances list (DSL) (Environment Canada 2012). A compiled list of all substances thought to be in commercial use in Canada (~ 24,000 chemicals). Thus, it includes chemicals with potential human or ecological exposure.
- The Endocrine Disruption Screening Program (EDSP) universe of ~ 10,000 chemicals. The U.S. EPA's EDSP is required to test certain chemicals for their potential for endocrine disruption (U.S. EPA 2014d).
- A list of ~ 15,000 chemicals used as training and test sets for the different models implemented in the U.S. EPA's Estimation Program Interface (EPI Suite™) to predict physico-chemical properties (U.S. EPA 2014e).

This virtual chemical library has undergone stringent chemical structure processing and normalization for use in the QSAR modeling study (see "Chemical Structure Curation") and made available for download on ToxCast™ Data web site

under CERAPP data (https://www3.epa.gov/research/COMPTOX/CERAPP_files.html, PredictionSet.zip) (U.S. EPA 2016), is intended to be employed for a large number of other QSAR modeling projects, not just those focused on endocrine-related targets.

Experimental evaluation set. A large volume of estrogen-related experimental data has accumulated in the literature over the past two decades. The information on the estrogenic activity of chemicals was mined and curated to serve as a validation set for predictions of the different models. For this purpose, *in vitro* experimental data were collected from different overlapping sources, including the U.S. EPA's HTS assays, online databases, and other data sets used by participants to train models:

- HTS data from Tox21 project consisting of ~ 8,000 chemicals evaluated in four assays (Attene-Ramos et al. 2013; Collins et al. 2008; Huang et al. 2014; Shukla et al. 2010; Tice et al. 2013), extending beyond the 1,677 used in the training set.
- The U.S. FDA Estrogenic Activity Database (EADB), which consists of literature derived ER data for ~ 8,000 chemicals (Shen et al. 2013).
- Estrogenic data for ~ 2,000 chemicals from the METI (Ministry of Economy, Trade and Industry, Japan) database (METI 2002).
- Estrogenic data for ~ 2,000 chemicals from ChEMBL database (Gaulton et al. 2012).

The full data set consisted of > 60,000 entries, including binding, agonist, and antagonist information for ~ 15,000 unique chemical structures. For the purpose of this project, this data set was cleaned and made more consistent by removing *in vivo* data, cytotoxicity information, and all ambiguous entries (missing values, undefined/nonstandard end points, and unclear units). Only 7,547 chemical structures from the experimental evaluation set that overlapped with the CERAPP prediction set, for a total of 44,641 entries, were kept and made available for download on the U.S. EPA ToxCast™ Data web site (https://www3.epa.gov/research/COMPTOX/CERAPP_files.html, EvaluationSet.zip) (U.S. EPA 2016). The non-CERAPP chemicals were excluded from the evaluation set (see "Chemical Structure Curation" section). Then, all data entries were categorized into three assay classes: (a) binding, (b) reporter gene/transactivation, or (c) cell proliferation. The training set end point to model is the ER model AUC that parallels the corresponding individual assay AC₅₀ values, and therefore all units for activities in the experimental data set were converted to μM to have approximately equivalent concentration–response values for the evaluation set. Chemicals with cell proliferation assays were considered as actives if they exceeded an arbitrary threshold of 125% proliferation. For entries where testing

concentrations were reported in the assay name field, those values were converted to μM and considered as the AC₅₀ value if the compound was reported as active. All inactive compounds were arbitrarily assigned an AC₅₀ value of 1 M.

Chemical Structure Curation

Chemical structures collected from different public sources contained many duplicates, and inconsistencies in the molecular structures. Hence, a structure curation process was carried out to derive a unique set of QSAR-ready structures. All participating groups then used this consistent set of structures for both training and prediction steps. It should be noted that each group likely employed different descriptor calculation software, which could effectively alter structures in some cases. Several different curation approaches were combined into a unique procedure used for this project (Fourches et al. 2010; Wedebye et al. 2013). The free and open-source data-mining environment KNIME (Konstanz Information Miner) was selected to design a curation workflow to process all structures and provide consistent training and prediction sets (Berthold et al. 2007). The workflow performed a series of curation steps:

- 1) The original files containing structures in different formats were parsed, checked for valences, and for the integrity of the required structural information to render the molecules. Invalid entries were corrected by retrieving a new structure from online databases using web services [PubChem (NIH 2015), ChemSpider (Royal Society of Chemistry 2015)] or removed if ambiguous.
- 2) The first filter was applied to check for the presence of carbon atoms and remove inorganic compounds.
- 3) The structures were desalted, and inorganic counterions were removed.
- 4) The second filter, based on molecular weight, was applied and chemicals exceeding a threshold of 1,000 g/mol were removed to speed up molecular descriptor calculations and model calibration.
- 5) Valid QSAR modeling practice requires all chemicals to be structurally consistent by converting tautomers to unique representations. Thus, a series of transformations was applied on the structures to standardize nitro and azide mesomers, keto-enol tautomers, enamine-imine tautomers, yno-ol-ketene, and other conversions (ChemAxon 2014; Reusch 2013; Sitzmann et al. 2010).
- 6) These transformations were followed by neutralizing the charged structures, when possible, and removing the stereochemistry information.
- 7) Explicit hydrogen atoms were added, and structures were aromatized according to

Hückel's rules implemented in KNIME (Berthold et al. 2007).

- 8) The duplicates were removed using the IUPAC (International Union of Pure and Applied Chemistry) InChI (International Chemical Identifier) codes because these are unequivocal identifiers.
- 9) The final filter was applied to remove chemicals containing metals that often cause problems in molecular descriptor calculations.

Both training and prediction sets were processed by the same structure curation workflow. At the end of this procedure, 32,464 unique structures—the 32 K set—remained in the prediction set and 1,677 in the training set. These two data sets are made available for download in structure data file (SDF) format on the U.S. EPA ToxCast™ Data web site (https://www3.epa.gov/research/COMPTOX/CERAPP_files.html, TrainingSet.zip and PredictionSet.zip) (U.S. EPA 2016). The identity of these chemicals (name, CASRN) was not provided to the participating modeling groups during the modeling process.

Modeling Approaches

The participant groups adopted different approaches and used several software programs (proprietary or open-source [commercial or free]) to calibrate categorical and continuous models to the training data (Table 1). A categorical model is one that provides an active/inactive call for each chemical, whereas a continuous model provides a prediction of the potency (in μM) for each active chemical. Models were developed using both well-known and innovative methods including partial least-squares (PLS) (Stähle and Wold 1987; Wold et al. 2001), partial least-squares discriminant analysis (PLS-DA) (Frank and Friedman 1993; Nouwen et al. 1997), decision forest (DF) (Hong et al. 2005, 2004; Tong et al. 2003; Xie et al. 2005), three-dimensional (3D) quantitative spectral data–activity relationship (QSDAR) (Beger et al. 2001; Beger and Wilkes 2001; Slavov et al. 2013), support vector machines (SVM) (Cristianini and Shawe-Taylor 2000), *k* nearest neighbors (kNN) (Cover and Hart 1967; Kowalski and Bender 1972), associative artificial neural networks (ASNN) (Tetko 2002a, 2002b), PASS algorithm derived from Naïve Bayes classifier (Poroikov et al. 2000), self-consistent regression with radial basis function interpolation (RBF-SCR) (Zakharov et al. 2014), OCHEM machine learning methods (Tetko et al. 2014), docking and *consensus* of different approaches (Horvath et al. 2014; Ng et al. 2014; Sushko et al. 2011). The set of 1,677 chemicals provided by the U.S. EPA was used by more than 90% of the participating groups as a training set to fit their models (Judson

et al. 2015), but some pre-existing models were also used that had been trained using other data sets from the literature such as METI (2002). In addition, each group performed its own analysis to select the appropriate chemicals to be considered as a training set according to their particular modeling procedure. For descriptor calculation and docking procedures, some of the programs used were LeadScope (Roberts et al. 2000), PaDEL-Descriptor (Yap 2011), QikProp (version 3.4, <http://www.schrodinger.com/QikProp/>), multilevel and quantitative neighborhoods of atoms (MNA, QNA) used by GUSAR and PASS (Filimonov et al. 2009; Poroikov et al. 2000), DRAGON (Talete srl 2012), Mold2 (Hong et al. 2008, 2012), GLIDE (version 6.5, <http://www.schrodinger.com/Glide/>), AutoDock (Goodsell et al. 1996), ISIDA (Varnek et al. 2008), and other fingerprint generators. Some of the participants applied feature selection techniques, such as genetic algorithms (GAs) (Davi 1991) and random forest (RF) (Breiman 2001). These techniques were applied after calculating descriptors to reduce collinearity and variable dimensionality to keep only the most informative descriptors in the models.

Evaluation Procedure for the Categorical and Continuous Models

All molecular structures of chemicals collected for the evaluation set from the different sources were curated and standardized using the previously described KNIME workflow (Table S2, step 2). All data used as the evaluation set for categorical and continuous models are available on the U.S. EPA ToxCast™ web site (https://www3.epa.gov/research/COMPTOX/CERAPP_files.html, EvaluationSet.zip) (U.S. EPA 2016).

Standard InChI codes were generated in KNIME and used to identify the chemicals. Data-mining tools available in the KNIME environment were used to concatenate and unify the different information fields from the different sources (CASRN, chemical name, original structure, standardized structure, InChI code, assay name, assay class, protein subtype, species, end point name, end point value, end point unit, and literature reference). Although ToxCast™ chemicals were used in the training sets of many models, they were not removed from the evaluation set to investigate how the predictions will perform on the literature data because there are differences between the AUC values and the literature data and because the sources from which the evaluation set was collected were not fully verified (we cannot assume that all cytotoxicity information was already fully cleaned).

Evaluation set for categorical models. An important issue with the literature-derived evaluation set was the inconsistency of the results from different sources. To minimize this, the available entries for each chemical structure were grouped into binders, agonists, and antagonists. The results were then categorized into active and inactive classes using all available literature sources by applying three rules:

- 1) If, for a specific chemical within one of the three classes (binding, agonist, and antagonist), the disagreement among the different sources exceeded 20% (e.g., two sources indicating active agonist and three indicating inactive agonist), that chemical was removed from the evaluation data set of that specific class.
- 2) If a chemical was an active agonist or antagonist, it also was considered as

an active binder if the information was not available.

- 3) If a chemical was an inactive agonist and inactive antagonist, it was considered also as nonbinder if the information was not available.

This procedure resulted in a total of 7,522 unique chemical structures with activity data to be used for evaluation of the categorical models (Table 2). It is also available for download on the U.S. EPA ToxCast™ web site (https://www3.epa.gov/research/COMPTOX/CERAPP_files.html, EvaluationSet.zip) (U.S. EPA 2016).

Evaluation set for continuous models. For active chemicals with available quantitative information from concentration-response assays, the log₁₀-median of the literature values was calculated. Only entries with equivalent end points were considered (e.g., PC50 and EC50). This resulted in 7,253 unique chemicals with quantitative information (Table 3 and https://www3.epa.gov/research/COMPTOX/CERAPP_files.html, EvaluationSet.zip) (U.S. EPA 2016). To reduce the variability that increased with the disparate literature sources, the chemicals with quantitative information were categorized into five potency activity classes: inactive, very weak, weak, moderate, and strong. These five classes were used to evaluate the quantitative predictions. A list of 36 known active and inactive reference chemicals was used for calibrating the mapping from quantitative potency values to the activity potency classes (Judson et al. 2015). These same chemicals were used to validate the mathematical model used to generate the AUC values for the training set. The following thresholds were applied to the concentration-response values:

Table 1. Methods adopted by the participant groups (alphabetic order) in the modeling procedure.

Model name	Calibration method	Descriptors software/type	Training set (No. of chemicals)	Predictions type
DTU	PLS/fragments	Leadscope	METI (595,481)/ToxCast™ (1,422)	Categorical
EPA_NCCT	GA + PLS-DA	PADEL	ToxCast™ (1,529)	Categorical
FDA_NCTR_DBB (Ng et al. 2014)	DF	Mold2	ToxCast™ (1,677)	Categorical
FDA_NCTR_DSB	PLS	3D-SDAR	ToxCast™ (1,019)	Categorical
ILS_EPA (Zang et al. 2013)	SVM + RF	Qikprop	ToxCast™ (1,677)	Categorical
IRCCS_CART (Roncaglioni et al. 2008)	CART-VEGA	2D descriptors	METI (806)	Categorical
IRCCS_Ruleset	Ruleset	SMARTS	ToxCast™ (1,529)	Categorical
JRC_Ispra (Poroikov et al. 2000)	PASS	MNA	—	Categorical
Lockheed Martin	kNN	Fingerprints	ToxCast™ (1,677)	Categorical + continuous
NIH_NCATS	Docking	AutoDock score	—	Categorical
NIH_NCI_GUSAR (Filimonov et al. 2009)	RBF-SCR	MNA, QNA	ToxCast™ (1,677)	Categorical
NIH_NCI_PASS (Poroikov et al. 2000)	PASS	MNA	ToxCast™ (1,677)	Categorical
OCHEM (2015)	Consensus	11 Descriptor types	ToxCast™ (1,660)	Categorical + continuous
RIFM	SVM	Fingerprints	ToxCast™ (1,677)	Categorical
Umeå (Rybacka et al. 2015)	ASNN	DRAGON	METI + (Kuiper et al. 1997; Taha et al. 2010)	Categorical
UNC_MML	SVM+RF	DRAGON	ToxCast™ (120)	Categorical
UNIBA (Trisciuzzi et al. 2015)	Docking	GLIDE score	ToxCast™ (1,677)	Categorical
UNIMIB	kNN	DRAGON + fingerprints	ToxCast™ (1,677)	Categorical
UNISTRA (Horvath et al. 2014)	SVM	ISIDA	ToxCast™ (1,529)	Categorical + continuous

Predictions type: A categorical model is one that provides an active/inactive call for each chemical, whereas a continuous model provides a prediction of the potency (in μ M) for each active chemical. Calibration methods: PLS (partial least-squares), PLS-DA (partial least-squares discriminant analysis), SVM (support vector machines), RF (random forest), DF (Decision forest), kNN (k nearest neighbors), ASNN (associative artificial neural networks), PASS (algorithm derived from Naïve Bayes classifier), RBF-SCR (self-consistent regression with radial basis function interpolation).

- Strong: Activity concentration below 0.09 μM .
- Moderate: Activity concentration between 0.09 and 0.18 μM .
- Weak: Activity concentration between 0.18 and 20 μM .
- Very Weak: Activity concentration between 20 and 800 μM .
- Inactive: Activity concentration higher than 800 μM .

The five classes were assigned scores from 0 (inactive) to 1 (strong) with 0.25 increments. Then, for each chemical, the arithmetic mean of the scores of the merged entries from different literature sources was calculated. A new class was assigned to the merged entries according to the following thresholds.

- Strong: Average score > 0.75
- Moderate: $0.5 < \text{Average score} \leq 0.75$
- Weak: $0.25 < \text{Average score} \leq 0.5$
- Very weak: $0 < \text{Average score} \leq 0.25$
- Inactive: Average score $= 0$

The number of entries in each class for binding, agonist, and antagonist are summarized in Table 3.

Evaluation procedure. This section is focused on the categorical models for their high number compared to the continuous models. The procedure used to evaluate the predictions of the participant groups was based on the categorical and continuous experimental data from ToxCast™ and the evaluation set from the literature. All continuous and categorical models for binding, agonist, and antagonist were evaluated separately on the overlap between their predicted chemicals and the following sets of chemicals (Table S3).

- Chemicals in the U.S. EPA's ToxCast™ data set ($n = 1,529$ chemicals after excluding those in the ambiguous AUC range of 0.01–0.1).
- All chemicals in the full literature data (all literature sources combined).
- All chemicals with at least two literature sources.
- All chemicals from the literature data excluding the very weak actives.
- Chemicals within the applicability domain (AD) of each model (if provided).
- Chemicals remaining after applying the previous three filters in steps 3, 4, and 5 to reduce ambiguous predictions (single

Table 2. Evaluation set for binary categorical models. Distribution of the number of active and inactive chemicals within the three different classes: binding, agonists and antagonists.

Class/activity	Active	Inactive	Total
Binding	1,982	5,301	7,283
Agonist	350	5,969	6,319
Antagonist	284	6,255	6,539
Total	2,017	7,024	7,522

The classification into actives and inactives is based on a consensus between the literature data sources that were in agreement.

literature source, very weak actives, and predictions outside the AD).

To evaluate the models on different criteria, we first determined the sensitivity (fraction of accurately predicted actives out of all actives), specificity (fraction of accurately predicted inactives out of all inactives), and balanced accuracy (BA; average of sensitivity and specificity) for each subgroup of chemicals according to each model. We then used BA values to derive two summary scores for each model, as described below.

Score 1. Evaluation includes BA of each of the six steps weighted by the fraction of predicted chemicals of the same step, as well as the fraction of the predicted chemicals out of the full prediction set. This score (Equation 1) favors models with a wider AD and those predicting a maximum number of chemicals.

$$\text{score}_1 = \frac{1}{3} \left(\frac{BA_{\text{ToxCast}} \times N_{\text{predToxCast}}}{N_{\text{ToxCast}}} + \frac{N_{\text{pred}}}{N_{\text{total}}} \right) + \frac{1}{N_{\text{filters}}} \sum_{i=1}^{N_{\text{filters}}} \left(\frac{BA_i \times N_{\text{pred}_i}}{N_{\text{total}_i}} \right) \quad [1]$$

where BA is balanced accuracy, N_{pred} is the number of predicted chemicals by a specific model, N_{total} is the total number of chemicals in the prediction set, N_{filters} represents the number of five filters applied to the evaluation set chemicals and i the steps 2, 3, 4, 5, and 6.

Score 2. Evaluation includes the BA of the model on the ToxCast™ data and the BA on the unambiguous chemicals (i.e., the subgroup of chemicals from the literature that remained after excluding chemicals with only one literature source, very weak chemicals, and chemicals outside of the AD, if provided). It favors models that focused on predicting more accurately but potentially with a narrower AD (Equation 2).

$$\text{score}_2 = \frac{1}{2} (BA_{\text{ToxCast}} + BA_{\text{all filters}}) \quad [2]$$

The quantitative predictions were evaluated as categorical models (using the BA) of the five classes after converting the numerical predictions to potency classes as defined earlier (see "Evaluation set for continuous models" section). Scores of the continuous models were calculated using Equation 2.

Table 3. Evaluation set for quantitative models. Distribution of the number of chemicals in the five potency levels within the three different classes (binding, agonists, and antagonists), classifications based on average scores.

Class/activity	Inactive	Very weak	Weak	Moderate	Strong	Total
Binding	5,042	685	894	72	77	6,770
Agonist	5,892	19	179	31	42	6,163
Antagonist	6,221	76	188	10	10	6,505
Total	6,892	702	916	81	93	7,253

The classification of the chemicals in the five potency levels is based on the concentration responses from the literature sources that were in agreement.

Consensus Modeling

The *consensus* predictions were generated for binders, agonists, and antagonists separately. For each chemical, we derived the average Score 2 value for all categorical models that predicted the chemical as active, and the average Score 2 value for all categorical models that predicted the chemical as inactive; we used the higher of the two averages to classify the chemical as active or inactive. Models that did not provide a prediction for the chemical in question were not included when deriving the average scores. We used Score 2 to derive the consensus classifications because its value for individual models is not penalized for the number of chemicals not predicted by the model. Also, the concordance among models on both active and inactive classes was calculated for each chemical as the fraction of models with positive and negative prediction, respectively.

Considering only the models that provided predictions, the sum of the concordance among models for actives and inactives is equal to 1. Because most models were associated with comparable scores, the average score used to classify chemicals was mostly in agreement with model concordance (i.e., the average score for actives is high when the concordance among the models with active predictions is high and vice versa). The few exceptions were noticed when model concordance was around 0.5, which means only one or two models were driving the classification.

For continuous predictions, the weight (w) for each chemical i was calculated from the scores (Equation 3):

$$w_i = \text{score}_i / \sum_{j=1}^n \text{score}_j \quad [3]$$

where n is the total number of models that provided predictions for the chemical i , and score_j is the score of the j th model predicting chemical i .

Next, the *consensus* potency level C_i of each chemical was determined using the predicted potency classes P_j of the n available models and their corresponding weights w as follows (Equation 4):

$$C_i = \sum_{j=1}^n w_j \times P_j \quad [4]$$

Mansouri et al.

Results and Discussion

Models and Evaluation

A total of 48 models were received from the 17 participant groups. Each group provided at least 1 categorical model for binding. Only 8 groups built models for agonists, and 6 groups built models for antagonists. The limited number of models for agonists and antagonists was the result of the low number of actives, which caused the training set to be highly unbalanced. The total number of models in each class (Table 1; see also Tables S3 and S5) was *a*) binding models: 21 categorical and 3 continuous, *b*) agonist models: 11 categorical and 3 continuous, and *c*) antagonist models: 8 categorical and 2 continuous.

The participating groups provided predictions for uneven fractions of the 32 k set. AD information on model predictions was provided by only six groups. All predictions for the individual models are provided on the U.S. EPA ToxCast™ web site (https://www3.epa.gov/research/COMPTOX/CERAPP_files.html, Models.zip) (U.S. EPA 2016).

The same evaluation procedure was applied to all models following the previously described steps. Note that some models were built using training sets other than what was provided in CERAPP and that these alternative training sets were not all publicly available. Hence, none of the training set chemicals were excluded from the evaluation sets (Table 1). Each model was evaluated on the overlap between the predicted chemicals and the two previously mentioned data sets: ToxCast™ data and the evaluation set collected from the literature. The evaluation results for categorical models are summarized in Table S3. The detailed statistics, including sensitivity and specificity, are provided in Table S4.

Most compounds were predicted as inactive and the models seemed to be more in agreement in predicting inactives than active compounds. Only 757 chemicals (2.33%) are predicted as actives by more than 75% of binding models. The agreement among the binding models for the 32 k set of the prediction set is illustrated in Figure S1.

Most categorical models (binding, agonist, and antagonist) are associated with high balanced accuracies on the ToxCast™ data (> 0.8), with no clear difference between models that used it as a training set and those that did not (see Table S3). However, for the evaluation set from the literature, the BA is clearly lower for all models (< 0.7). Nonetheless, the BA increased after removing chemicals with only one source from the literature data. This result could mean that this first filter (i.e., removing chemicals with limited information in the literature for

being either positive or negative) reduced the uncertainty in the experimental data from the literature. This is in agreement with related studies showing that the results of QSAR models may change depending on the robustness of the experimental values (Steinmetz et al. 2014). The second filter (i.e., removing very weak actives) also increased the BA, which suggests that the literature data may contain a number of false positives. Alternatively, the *in vitro* assays used by ToxCast™/Tox21 only test chemicals up to 100 μM, so very weak chemicals may not be picked up by these assays and some of the literature reports may have tested chemicals up to much higher concentrations.

Finally, removing predictions outside the AD did not show improvement of the BA of the categorical models (see Table S3). This is in agreement with literature sources showing that predictions outside the AD are not always less accurate than those within its limits (Sahigara et al. 2012). The performance of most models showed a clear improvement of 0.05 to 0.1 on the BA after applying all the filters on the literature data to keep only the unambiguous chemicals. We believe that this effectively reduced the uncertainty of the literature sources. This step also highlighted differences between ToxCast™ and the literature data and confirmed the existence of uncertainty in the literature data. Uncertainty and data discordance was also reported in literature review of *in vivo* uterotrophic bioassays (Kleinstreuer et al. 2015).

The calculated scores for categorical models (see Table S3) take into consideration the whole prediction set (Score_1) and the accuracy of the model on its most reliable predictions (Score_2). The models that provided predictions for the whole or most of the 32 k set of chemicals, and had wide ADs, showed high Score_1 values (Umeå 0.82, OCHEM 0.83). Whereas models with predictions for smaller fractions of the prediction set and narrow AD showed better Score_2 values (UNIMIB_2 0.85, UNIBA 0.80). NIH_NCI_GUSAR (0.87

and 0.84) and FDA_NCTR_DBB (0.88 and 0.84) showed the highest values for both Score_1 and Score_2. Part of the differences among model scores could result from the uncertainty in the literature data.

The BAs of all antagonist models was low compared with binding and agonist models (see Table S3). This may be due to the highly unbalanced training set with a low number of active antagonist chemicals. Additionally, antagonism activity (in either ToxCast™ or the literature) can be confounded with cytotoxicity because antagonist transactivation assays are loss-of-signal assays.

The predictions of all continuous models were first converted to five classes using the list of reference chemicals as described in the evaluation set section (see "Evaluation set for continuous models" section). The predictions were then evaluated on the ToxCast™ data and the literature data to calculate the average of BA of the different evaluation steps as the score of each model (see Table S5). All models showed high BA on ToxCast™ data and relatively good BA on the evaluation set.

Consensus Model

The *consensus* predictions were first evaluated on the ToxCast™ data and then on the evaluation set from the literature. The total number of predicted active binders was 2,661 out of the 32 k set of chemicals (8.2%) based on the method described in the "Materials and Methods" section "Consensus Modeling."

Confusion matrices (Table 4) and prediction statistics (Table 5) revealed a clear accuracy difference between the categorical *consensus* for binding on the ToxCast™ data and on the evaluation set. This difference could result from the fact that the ToxCast™ data, based on a model with inputs from 18 different assays, were used by most of the models as a training set, which we presume reduces the uncertainty. This is in contrast to the literature data, where the number of sources per chemical varied from one to a few hundreds. When only the subset of the evaluation set with more than six literature sources

Table 4. Confusion matrices of categorical *consensus* predictions for binding.

Observed/predicted	ToxCast™ data predicted actives	ToxCast™ data predicted inactives	Literature evaluation set (all: 7,283) predicted actives	Literature evaluation set (all: 7,283) predicted inactives
Observed actives	76	13	467	1,515
Observed inactives	25	1,415	268	5,033

Table 5. Statistics of categorical *consensus* predictions for binding on ToxCast™ and literature data.

Statistics/used data	ToxCast™ data	Literature evaluation set (all: 7,283)	Literature evaluation set (> 6 sources: 1,257)
Sensitivity	0.85	0.23	0.85
Specificity	0.98	0.95	0.97
Balanced accuracy	0.92	0.59	0.91

The literature data with more than six sources represents the most consistent part of the evaluation set.

per chemical was considered, a large increase in the sensitivity was noticed (0.23 to 0.85).

To better understand the effect of the number of sources on the classification accuracy, ROC (receiver operating characteristic) curves were made using the fraction of the binding models in each class as a threshold for the classification predictions and increasing the number of literature sources of the evaluation set. The ROC plot shows an improvement of the classification accuracy of the *consensus* model as the number of sources increases (Figure 1). Note that the same level of consistency (i.e., 80%) was required to merge the sources regardless of the number of sources (see rule 1 in the "Evaluation set for categorical models" section). This could lead to the conclusion that the low classification accuracy on the full literature data is not because of a lack of accuracy of the *consensus* predictions, but rather to noise and experimental uncertainty in the literature data. We assume that the high number of false negatives in the confusion matrix of Table 4 is caused by false positives in the full literature data for chemicals tested only a small number of times. Thus, by considering a higher number of sources (i.e., six), the number of false positives is reduced from the evaluation set and so the number of predicted false negatives decreased. This is in agreement with what was observed in the literature (Steinmetz et al. 2014).

Corrections to the Consensus Model

The first step of *consensus* modeling was conducted in an independent way for the categorical and continuous models on binding, agonist, and antagonist predictions. This led to a number of inconsistencies because some chemicals were predicted as active in categorical predictions but inactive in quantitative and vice versa. In addition, some chemicals were predicted as active agonists or antagonists but non-binders. To make all predictions more consistent, a number of corrections were applied on the first *consensus* predictions. Because the goal of this project was to help in a regulatory prioritization procedure, the modifications aimed to reduce the number of false negatives but without adding an excess of false positives. The rules that were followed to obtain the final *consensus* predictions are as follows:

- 1) If a chemical *i* is active in the categorical *consensus*, then it is also considered active in the quantitative *consensus*.
- 2) If a chemical *i* is active in the quantitative *consensus* and predicted as active by at least three categorical models, then it is also considered active in the categorical *consensus*.
- 3) If a chemical *i* is predicted active by less than three categorical models, then it is considered inactive also in quantitative *consensus*.

These three rules were applied on the agonist and antagonist *consensus* models first, then on the binding *consensus*. A fourth rule was added to establish consistency between agonist and antagonist *consensus* models and the binding *consensus* model.

- 4) If a chemical *i* is an active agonist or active antagonist, then it is considered as active in categorical binding *consensus*, and its potency level in the quantitative binding *consensus* is made equal to its potency level as agonist/antagonist.

An analysis of variance in concordance in each potency level of the active chemicals in the continuous models (very weak, weak, moderate, and strong) is presented as a box-plot in Figure 2. Based on this figure, we noticed a correlation between the concordance of the categorical models and the potency level of active chemicals. This implies that models are more in agreement for strong actives and that the weaker a chemical is the more difficult it is to accurately predict. Therefore, the very weak chemicals are the main source of discordance among the different *in silico* models and also are the most uncertain experimentally. This relationship between positive concordance (agreement between models on predictions for active chemicals) and potency level for active chemicals can be used to set a quantitative prediction to the newly reclassified active chemicals using the previously mentioned rule 1 of the corrections applied to the *consensus* predictions. The following thresholds were considered for each potency level:

- Strong: Concordance among models ≥ 0.9
- Moderate: $0.75 \leq$ Concordance among models < 0.9
- Weak: $0.6 \leq$ Concordance among models < 0.75
- Very weak: Concordance among models < 0.6

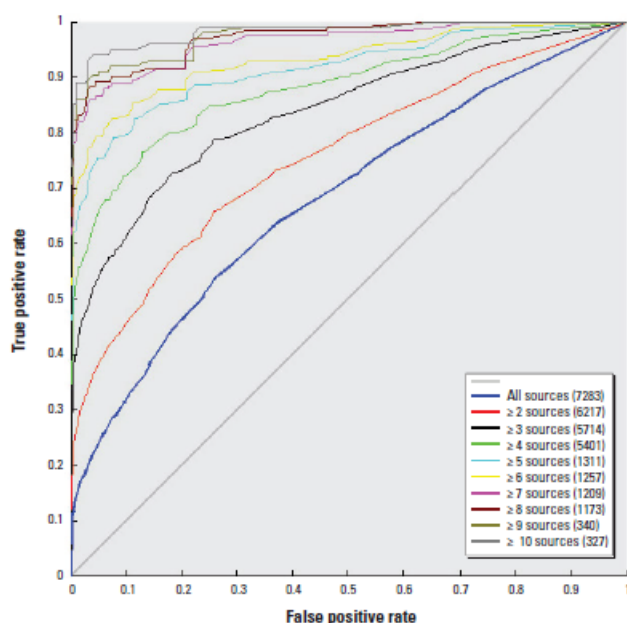


Figure 1. ROC curves of the categorical corrected consensus predictions for binding evaluated against different sets of the evaluation set with variable numbers of literature sources. The number of available chemicals in the evaluation set (between brackets) decreased with higher numbers of literature sources. The true and false positive rates are determined based on the number of actives in the different sets of the evaluation set.

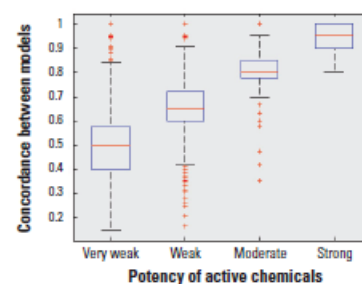


Figure 2. Box-plot of the positive class potency levels in the corrected quantitative *consensus* predictions for binding. The concordance between models is the fraction of the number of models that agrees on the prediction of a certain chemical. Boxes extend from the 25th to the 75th percentile, horizontal bars represent the median, whiskers indicate the 10th and 90th percentiles, and outliers are represented as points.

After applying the four correction rules on consensus predictions, the total number of chemicals predicted as actives increased from 2,661 to 4,001, which corresponds to 12.3% of the total number of the prediction set (32,464). Table 6 shows the number of reclassified chemicals based on each one of the four correction rules applied to the consensus predictions. After this step, the predicted activity of several chemicals has changed. The structural information of chemicals and the predictions of the *consensus* model for the whole 32 k set are provided on the U.S. EPA ToxCast™ web site (see https://www3.epa.gov/research/COMPTOX/CERAPP_files.html, PredictionSet.zip) (U.S. EPA 2016).

The confusion matrices and statistics for the binding categorical *consensus* model after modifications evaluated on ToxCast™ data and the literature data are presented in Table 7 and Table 8, respectively. The effect of the number of sources on the classification accuracy of the *consensus* model is illustrated by a bar plot in Figure S2. This figure shows an improvement of sensitivity with the increase in the number of literature sources in the evaluation set (from ~ 0.3 with at least one source to > 0.6 with six sources and more). This is translated into an increase in BA, whereas specificity is almost constant (~ 0.9) because of the high number of

inactives compared to active compounds.

The results of this project and the ToxCast™ data used as the training set are published online in the EDSP21 dashboard, together with other structural and experimental assay information (see “Consensus CERAPP QSAR ER Model Predictions” under “Chemical Summary” tab on <http://actor.epa.gov/edsp21>) (U.S. EPA 2014c). A comparison of the single classification models to the *consensus* predictions for the whole 32 k set of chemicals is provided in Table S6. The calculations are done using the categorical consensus predictions as the “observed response.”

For regulatory or prioritization purposes, one could use a looser definition of active (i.e., allow more disagreement among models) in order to further reduce the chance of false negatives. Figure 3 shows the number of chemicals that can be predicted as potential actives by the categorical consensus for binding using various positive concordance (agreement on actives between the included models) thresholds. When this threshold is set to 0.2, an additional 6,742 more chemicals can be added to the potential positives (this refers to the available binding models). This figure also shows the BA variations at different numbers of literature sources in the literature. Balanced accuracy increases as the concordance threshold increases from 0 to 0.2

because sensitivity increases (false negatives decrease) as the number of chemicals classified as active increases. For chemicals with the highest data quality (seven or more sources), the BA curve reaches a plateau at concordance thresholds of 0.4–0.5, and the number of chemicals classified as active is consistent with the number of active chemicals predicted from our consensus model ($n = 4,001$). However, higher concordance thresholds result in declining BA due to increasing numbers of false positive predictions (i.e., decreasing specificity).

Conclusion

The collaborative efforts of the CERAPP participants resulted in *consensus* predictions of the ability of chemicals to interact with ER. Up to 48 separately developed categorical and continuous models were received from 17 research groups from the United States and Europe. Separate models were built for agonist, antagonist, and binding activity. The models were applied to a large collection of 32,464 chemical structures that approximate the human exposure universe (chemicals with potential human exposure). A KNIME

Table 6. Number of chemicals reclassified after applying each one of the four prediction correction rules.

Rule used for each class	Rule 1			Rule 2			Rule 3			Rule 4	
	Agonist	Antagonist	Binding	Agonist	Antagonist	Binding	Agonist	Antagonist	Binding	Binding	
Number of chemicals	1,288	2,760	1,587	217	14	344	145	161	38	966	

Rule 1: Chemicals that changed from inactive to active in the quantitative consensus based on the categorical consensus. Rule 2: Chemicals that changed from inactive to active in the categorical consensus based on the quantitative consensus. Rule 3: Chemicals that changed from active to inactive in the quantitative consensus based on the predictions of the categorical consensus. Rule 4: Chemicals that changed from inactive to active in the categorical binding consensus based on their agonist and antagonist activity in the categorical consensus.

Table 7. Confusion matrices of the modified categorical *consensus* predictions for binding.

Observed/predicted	ToxCast™ data		Literature evaluation set (All: 7,283)	
	predicted actives	predicted inactives	predicted actives	predicted inactives
Observed actives	83	6	597	1,385
Observed inactives	40	1,400	463	4,838

Table 8. Statistics of the modified categorical *consensus* for binding predictions on ToxCast™ and literature data.

Statistics/used data	ToxCast™ data	Literature evaluation set (All: 7,283)	Literature evaluation set (> 6 Sources: 1,275)
Sensitivity	0.93	0.30	0.87
Specificity	0.97	0.91	0.94
Balanced accuracy	0.95	0.61	0.91

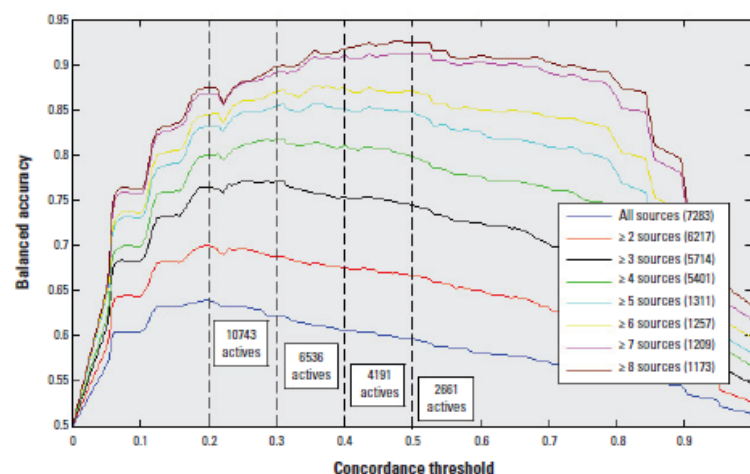


Figure 3. Variation of the balanced accuracy of the corrected categorical consensus predictions for binding with positive concordance (agreement between models on predictions for active chemicals) threshold at different numbers of literature sources.

workflow was developed to carefully curate the large collection of chemical structures to ensure consistency in model development and evaluation. Most of the models were trained using activities derived from a data set combining 18 *in vitro* assays from ToxCast™ probing various points of the ER pathway. Models were then evaluated using the ToxCast™ data plus a collection of ER *in vitro* data from the literature. After this process, categorical predictions were combined into a consensus to classify the chemicals into actives and inactive, while continuous predictions were combined to classify the actives into 4 different potency classes: very weak, weak, moderate, and strong.

One major observation was that most models had comparable performances, independent of the methods used, with a slight improvement for models with narrow ADs. A second and, perhaps, more important observation is that the most concordant predictions come from comparing the *consensus* of many models with a *consensus* of many literature sources. For instance, when comparing the *consensus* of the categorical binding models with the evaluation set from the literature for chemicals with seven or more sources, we achieve a balanced accuracy of about 90% (Table 8).

We propose several important conclusions from our results. First, there does not appear to be an optimal modeling approach (combination of descriptor set, feature selection, or machine learning algorithm) that will solve the QSAR/docking problem and achieve perfect prediction accuracies. Second, there are inherent limitations to the accuracy of the data being used to train QSAR and docking models. Our analysis of the literature data showed a disagreement in the reported activity of many chemicals. The sources of discrepancy include limits to the concentration ranges tested, true differential activity among tissue sources [e.g., the presence of selective ER modulators, SERMs (selective estrogen receptor modulators)], and a variety of experimental artifacts and errors. Figure 2 shows that the most consistent predictions are achieved for the most potent compounds, whereas weaker compounds are called inactive by some laboratories because these compounds were not tested at a high enough concentration. So chemicals with very weak activity would be more likely to be incorrectly classified as inactive than more potent chemicals. Therefore, 100% accuracy cannot be achieved due to these limitations in the experimental data used for training and evaluation. Figures 1 and 3 help to illustrate this point by showing that higher consistency in the experimental data is associated with an increase in the concordance among model predictions. But this comes at the cost of excluding parts

of the experimental data. So, just as every model has limitations, every *in vitro* assay also has inherent variability in its results.

The major purpose of this study was to identify potential ER actives out of the large universe of chemicals to which humans potentially are exposed using a *consensus* of *in silico* models to overcome the limitations of single models. Most of the chemicals in this collection were predicted to be negatives, with a high agreement among the individual models. The disagreement was the highest for chemicals with weak activity (Figure 2). This disagreement is driven by the difficulties in experimentally assessing the activity of these weak chemicals. In total, the consensus predicted 4,001 chemicals as actives. The testing of these active chemicals will be prioritized from the most potent to the least according to the continuous model *consensus* predictions. There are 6,742 more chemicals that 20–50% of the models predicted to be positive, which could also be candidates for follow-up analyses. Although this large number of chemicals (~10,000 in total) appears to be a daunting set to evaluate experimentally, this is equivalent in size to the current Tox21 library already being tested for activity in ER and many other targets.

In summary, this project demonstrates the feasibility of screening a large and toxicologically relevant library of chemical structures in an extensive battery of QSAR and docking models to meet important goals in human and environmental health. ER provides a good initial case because of the ready availability of experimental data and pre-existing models. However, through the ToxCast™ and Tox21 programs, and through other large scale data-integration projects, equivalently large data sets will become available for other multiple targets of environmental importance.

REFERENCES

- Adler S, Basketter D, Croton S, Pelkonen O, van Benthem J, Zuang V, et al. 2011. Alternative (non-animal) methods for cosmetics testing: current status and future prospects—2010. *Arch Toxicol* 85:367–485.
- Attene-Ramos MS, Miller N, Huang R, Michael S, Itkin M, Kavlock RJ, et al. 2013. The Tox21 robotic platform for the assessment of environmental chemicals—from vision to reality. *Drug Discov Today* 18:716–723. doi: 10.1016/j.drudis.2013.05.015.
- Beger RD, Buzatu DA, Wilkes JG, Lay JO Jr. 2001. ¹³C NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroids binding the aromatase enzyme. *J Chem Inf Comput Sci* 41:1360–1366.
- Beger RD, Wilkes JG. 2001. Developing ¹³C NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin. *J Comput Aided Mol Des* 15:659–669.
- Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meini T, et al. 2007. KNIME: the Konstanz Information Miner. In: *Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V.*, Albert-Ludwigs-Universität Freiburg, 7–9 March 2007, Heidelberg, Germany. Studies in Classification, Data Analysis, and Knowledge Organization (Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, eds). Heidelberg, Germany:Springer, 319–326.
- Birnbaum LS, Fenton SE. 2003. Cancer and developmental exposure to endocrine disruptors. *Environ Health Perspect* 111:389–394. doi: 10.1289/ehp.5686.
- Breiman L. 2001. Random forests. *Mach Learn* 45:5–32.
- ChemAxon. 2014. Standardizer. Structure Canonicalization and More. Available: <http://www.chemaxon.com/products/standardizer/> [accessed 26 November 2014].
- Cohen Hubal EA, Richard A, Aylward L, Edwards S, Gallagher J, Goldsmith MR, et al. 2010. Advancing exposure characterization for chemical evaluation and risk assessment. *J Toxicol Environ Health B Crit Rev* 13:299–313.
- Colborn T, vom Saal FS, Soto AM. 1993. Developmental effects of endocrine-disrupting chemicals in wildlife and humans. *Environ Health Perspect* 101:378–384.
- Collins FS, Gray GM, Bucher JR. 2008. Toxicology. Transforming environmental health protection. *Science* 319:906–907.
- Cover T, Hart P. 1967. Nearest neighbor pattern classification. *IEEE Trans Inf Theory* 13:21–27.
- Cristianini N, Shawe-Taylor J. 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. 1st ed. New York, NY:Cambridge University Press.
- Davi L. 1991. Handbook of Genetic Algorithm. New York, NY:Van Nostrand Reinhold.
- Davis DL, Bradlow HL, Wolff M, Woodruff T, Hoel DG, Anton-Culver H. 1993. Medical hypothesis: xenoestrogens as preventable causes of breast cancer. *Environ Health Perspect* 101:372–377.
- Dearden JC, Cronin MTD, Kaiser KLE. 2009. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR QSAR Environ Res* 20:241–266.
- Diamanti-Kandarakis E, Bourguignon JP, Giudice LC, Hauser R, Prins GS, Soto AM, et al. 2009. Endocrine-disrupting chemicals: an Endocrine Society scientific statement. *Endocr Rev* 30:293–342.
- Dionisio KL, Frame AM, Goldsmith MR, Wambaugh JF, Liddell A, Cathey T, et al. 2015. Exploring consumer exposure pathways and patterns of use for chemicals in the environment. *Toxicol Rep* 2:228–237. doi: 10.1016/j.toxrep.2014.12.009.
- Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. 2007. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* 95:5–12.
- Egeghy PP, Judson R, Gangwal S, Mosher S, Smith D, Vail J, et al. 2012. The exposure data landscape for manufactured chemicals. *Sci Total Environ* 414:159–166.
- Environment Canada. 2012. Domestic Substances List. Available: <http://www.ec.gc.ca/lcpe-cepa/default.asp?lang=En&n=5F213FA8-1> [accessed 4 November 2012].
- FDA (U.S. Food and Drug Administration). 1996. Compilation of Laws Enforced by the U.S. Food and Drug Administration and Related Statutes. Washington, DC:FDA.
- Filimonov DA, Zakharov AV, Lagunin AA, Poroikov VV. 2009. QNA-based ‘Star Track’ QSAR approach. *SAR QSAR Environ Res* 20:679–709.
- Fourches D, Muratov E, Tropsha A. 2010. Trust, but

- verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* 50:1189–1204.
- Frank IE, Friedman JH. 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35:109–135.
- Fujita T, Iwasa J, Hansch C. 1964. A new substituent constant, π , derived from partition coefficients. *J Am Chem Soc* 86:5175–5180.
- Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40 (database issue):D1100–D1107.
- Goodsell DS, Morris GM, Olson AJ. 1996. Automated docking of flexible ligands: applications of AutoDock. *J Mol Recognit* 9:1–5.
- Hansch C, Deutsch EW. 1966. The structure–activity relationship in amides inhibiting photosynthesis. *Bibl Laeger* 112:381–391.
- Hansch C, Maloney PP, Fujita T, Muir RM. 1962. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194:178–180.
- Hilleman B. 1994. Environmental estrogens linked to reproductive abnormalities, cancer. *Chem Eng News* 72:19–23.
- Hong H, Slavov S, Ge W, Qian F, Su Z, Fang H, et al. 2012. Mold² molecular descriptors for QSAR. In: *Statistical Modelling of Molecular Descriptors in QSAR/QSPR* (Dehmer M, Varmuza K, Bonchev D, eds). Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 65–109.
- Hong H, Tong W, Perkins R, Fang H, Xie Q, Shi L. 2004. Multiclass Decision Forest—a novel pattern recognition method for multiclass classification in microarray data analysis. *DNA Cell Biol* 23:685–694.
- Hong H, Tong W, Xie Q, Fang H, Perkins R. 2005. An *in silico* ensemble method for lead discovery: decision forest. *SAR QSAR Environ Res* 16:339–347.
- Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, et al. 2008. Mold² molecular descriptors from 2D structures for cheminformatics and toxicoinformatics. *J Chem Inf Model* 48:1337–1344.
- Horvath D, Brown JB, Marcou G, Varnek A. 2014. An evolutionary optimizer of *libsvm* models. *Challenges* 5:450–472.
- Huang R, Sakamuru S, Martin MT, Reif DM, Judson RS, Houck KA, et al. 2014. Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Sci Rep* 4:5664, doi: 10.1038/srep05664.
- Jacobs M, Janssens W, Bernauer U, Brandon E, Coecke S, Combes R, et al. 2008. The use of metabolising systems for *in vitro* testing of endocrine disruptors. *Curr Drug Metab* 9:796–826.
- Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, et al. 2010. *In vitro* screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ Health Perspect* 118:485–492, doi: 10.1289/ehp.0901392.
- Judson RS, Kavlock RJ, Setzer RW, Hubal EAC, Martin MT, Knudsen TB, et al. 2011. Estimating toxicity-related biological pathway altering doses for high-throughput chemical risk assessment. *Chem Res Toxicol* 24:451–462.
- Judson RS, Magpantay FM, Chickarmane V, Haskell C, Tania N, Taylor J, et al. 2015. Integrated model of chemical perturbations of a biological pathway using 18 *in vitro* high throughput screening assays for the estrogen receptor. *Toxicol Sci* 148:137–154, doi: 10.1093/toxsci/kfv168.
- Judson RS, Martin MT, Egeghy P, Gangwal S, Reif DM, Kothiyia P, et al. 2012. Aggregating data for computational toxicology applications: the U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) system. *Int J Mol Sci* 13:1805–1831.
- Judson R, Richard A, Dix D, Houck K, Elloumi F, Martin M, et al. 2008. ACToR—Aggregated Computational Toxicology Resource. *Toxicol Appl Pharmacol* 233:7–13.
- Judson R, Richard A, Dix DJ, Houck K, Martin M, Kavlock R, et al. 2009. The toxicity data landscape for environmental chemicals. *Environ Health Perspect* 117:685–695, doi: 10.1289/ehp.0800168.
- Kavlock R, Dix D. 2010. Computational toxicology as implemented by the U.S. EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. *J Toxicol Environ Health B Crit Rev* 13:197–217.
- Kavlock RJ, Daston GP, DeRosa C, Fenner-Crisp P, Gray LE, Kaattari S, et al. 1996. Research needs for the risk assessment of health and environmental effects of endocrine disruptors: a report of the U.S. EPA-sponsored workshop. *Environ Health Perspect* 104:715–740.
- Kleinstreuer NC, Ceger PC, Allen DG, Strickland J, Chang X, Hamm JT, et al. 2015. A curated database of rodent uterotrophic bioactivity. *Environ Health Perspect* 124(5):556–562.
- Kowalski BR, Bender CF. 1972. The K-Nearest Neighbor Classification Rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Anal Chem* 44:1405–1411.
- Kuiper GG, Carlsson B, Grandien K, Enmark E, Haggblad J, Nilsson S, et al. 1997. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors α and β . *Endocrinology* 138:863–870.
- Mahoney MM, Padmanabhan V. 2010. Developmental programming: impact of fetal exposure to endocrine disrupting chemicals on gonadotropin-releasing hormone and estrogen receptor mRNA in sheep hypothalamus. *Toxicol Appl Pharmacol* 247:98–104.
- Martin MT, Dix DJ, Judson RS, Kavlock RJ, Reif DM, Richard AM, et al. 2010. Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program. *Chem Res Toxicol* 23:578–590.
- METI (Ministry of Economy Trade and Industry, Japan). 2002. Current Status of Testing Methods Development for Endocrine Disruptors. 6th Meeting of the Task Force on Endocrine Disruptors Testing and Assessment (EDTA). 24–2 June 2002. Tokyo, Japan. Available: <http://www.meti.go.jp/english/report/data/gEndoctexta.pdf> [accessed 3 July 2015].
- Mueller SO, Korach KS. 2001. Estrogen receptors and endocrine diseases: lessons from estrogen receptor knockout mice. *Curr Opin Pharmacol* 1:613–619.
- Muster W, Breidenbach A, Fischer H, Kirchner S, Müller L, Pähler A. 2008. Computational toxicology in drug development. *Drug Discov Today* 13:303–310.
- Ng HW, Zhang W, Shu M, Luo H, Ge W, Perkins R, et al. 2014. Competitive molecular docking approach for predicting estrogen receptor subtype α agonists and antagonists. *BMC Bioinformatics* 15(suppl 1):S4, doi: 10.1186/1471-2105-15-S11-S4.
- NIH (National Institutes of Health). 2015. The PubChem Database. Available: <http://pubchem.ncbi.nlm.nih.gov/> [accessed 26 January 2015].
- Nouwen J, Lindgren F, Hansen B, Karcher W, Verhaar HJM, Hermens JLM. 1997. Classification of environmentally occurring chemicals using structural fragments and PLS discriminant analysis. *Environ Sci Technol* 31:2313–2318.
- OCHEM (Online Chemical Database with Modeling Environment). 2015. CERAPP Models. Available: <https://ochem.eu/article/71005> [accessed 12 January 2015].
- Poroikov VV, Filimonov DA, Borodina YV, Lagunin AA, Kos A. 2000. Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. *J Chem Inf Comput Sci* 40:1349–1355.
- Reusch W. 2013. Reaction Examples. Examples of Organic Reactions. Available: <http://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/react2.htm> [accessed 25 November 2014].
- Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE Jr. 2000. LeadScope: software for exploring large sets of screening data. *J Chem Inf Comput Sci* 40:1302–1314.
- Roncaglioni A, Piclin N, Pintore M, Benfenati E. 2008. Binary classification models for endocrine disruptor effects mediated through the estrogen receptor. *SAR QSAR Environ Res* 19:697–733.
- Rotroff DM, Dix DJ, Houck KA, Knudsen TB, Martin MT, McLaurin KW, et al. 2013. Using *in vitro* high throughput screening assays to identify potential endocrine-disrupting chemicals. *Environ Health Perspect* 121:7–14, doi: 10.1289/ehp.1205065.
- Royal Society of Chemistry. 2015. ChemSpider Webservices. Available: <http://www.chemspider.com/AboutServices.aspx> [accessed 28 January 2015].
- Rybacka A, Rudén C, Tetko IV, Andersson PL. 2015. Identifying potential endocrine disruptors among industrial chemicals and their metabolites—development and evaluation of *in silico* tools. *Chemosphere* 139:372–378, doi: 10.1016/j.chemosphere.2015.07.036.
- Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. 2012. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 17:4791–4810.
- Shanley EK, Xu W. 2011. Endocrine disrupting chemicals targeting estrogen receptor signaling: identification and mechanisms of action. *Chem Res Toxicol* 24:6–19.
- Shen J, Xu L, Fang H, Richard AM, Bray JD, Judson RS, et al. 2013. EADB: an estrogenic activity database for assessing potential endocrine activity. *Toxicol Sci* 135:277–291.
- Shukla SJ, Huang R, Austin CP, Xia M. 2010. The future of toxicity testing: a focus on *in vitro* methods using a quantitative high-throughput screening platform. *Drug Discov Today* 15:997–1007.
- Sitzmann M, Ihlenfeldt WD, Nicklaus MC. 2010. Tautomerism in large databases. *J Comput Aided Mol Des* 24:521–551.
- Slavov SH, Pearce BA, Buzatu DA, Wilkes JG, Beger RD. 2013. Complementary PLS and KNN algorithms for improved 3D-QSAR consensus modeling of AhR binding. *J Cheminform* 5:47, doi: 10.1186/1758-2946-5-47.
- Stähle L, Wold S. 1987. Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *J Chemom* 1:185–196.
- Steinmetz FP, Enoch SJ, Madden JC, Nelms MD, Rodriguez-Sanchez N, Rowe PH, et al. 2014. Methods for assigning confidence to toxicity data with multiple values—identifying experimental outliers. *Sci Total Environ* 482–483:358–365.
- Sung E, Turan N, Ho PWL, Ho SL, Jarratt PDB, Waring RH, et al. 2012. Detection of endocrine disruptors—from simple assays to whole genome scanning. *Int J Androl* 35:407–414.
- Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, et al. 2011. Online Chemical

- Modeling Environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* 25:533–554, doi: 10.1007/s10822-011-9440-2.
- Taha MO, Tarairah M, Zalloum H, Abu-Sheikha G. 2010. Pharmacophore and QSAR modeling of estrogen receptor β ligands and subsequent validation and *in silico* search for new hits. *J Mol Graph Model* 28:383–400.
- Taletto srl. 2012. DRAGON (Software for Molecular Descriptor Calculations). Milano, Italy:Taletto srl.
- Tetko IV. 2002a. Associative neural network. *Neural Processing Letters* 16:187–199.
- Tetko IV. 2002b. Neural network studies. 4. Introduction to associative neural networks. *J Chem Inf Comput Sci* 42:717–728.
- Tetko IV, Sushko Y, Novotarskyi S, Patiny L, Kondratov I, Petrenko AE, et al. 2014. How accurately can we predict the melting points of drug-like compounds? *J Chem Inf Model* 54:3320–3329, doi: 10.1021/ci5005288.
- Tice RR, Austin CP, Kavlock RJ, Bucher JR. 2013. Improving the human hazard characterization of chemicals: a Tox21 update. *Environ Health Perspect* 121:756–765, doi: 10.1289/ehp.1205784.
- Tong W, Hong H, Fang H, Xie Q, Perkins R. 2003. Decision forest: combining the predictions of multiple independent decision tree models. *J Chem Inf Comput Sci* 43:525–531.
- Trisciuzzi D, Alberga D, Mansouri K, Judson R, Cellamare S, Catto M, et al. 2015. Docking-based classification models for exploratory toxicology studies on high-quality estrogenic experimental data. *Future Med Chem* 7:1921–1936, doi: 10.4155/FMC.15.103.
- U.S. EPA (U.S. Environmental Protection Agency). 1996. Drinking Water Contaminants – Standards and Regulations. Available: <http://water.epa.gov/lawsregs/guidance/sdwa/theme.cfm> [accessed 25 November 2014].
- U.S. EPA. 2014a. CPCat: Chemical and Product Categories. Exploring consumer exposure pathways and patterns of use for chemicals in the environment. *Toxicology Reports* 2:228–237. Curated chemical and product categories data were retrieved from the CPCat Database, U.S. EPA, RTP, NC. Available: <http://actor.epa.gov/cpcat> [accessed 26 November 2014].
- U.S. EPA. 2014b. Distributed Structure–Searchable Toxicity (DSSTox). Available: <http://www.epa.gov/nccit/dsstox/> [accessed 26 November 2014].
- U.S. EPA. 2014c. EDSP21 Dashboard. Endocrine Disruptor Screening Program for the 21st Century. Available: <http://actor.epa.gov/edsp21/> [accessed 12 January 2015].
- U.S. EPA. 2014d. Endocrine Disruption. Endocrine Disruptor Screening Program (EDSP). Available: <http://www.epa.gov/endo/#universe> [accessed 12 January 2015].
- U.S. EPA. 2014e. EPI Suite Data. Available: <http://esc.syrres.com/interkow/EpiSuiteData.htm> [accessed 26 April 2014].
- U.S. EPA. 2015. Chemicals under the Toxic Substances Control Act (TSCA). Available: <http://www.epa.gov/oppt/> [accessed 25 November 2014].
- U.S. EPA. 2016. Collaborative Estrogen Receptor Activity Prediction Project Data. Available: https://www3.epa.gov/research/COMPTOX/CERAPP_files.html [accessed 2 February 2016].
- UNEP, WHO (United Nations Environmental Programme, World Health Organization). 2013. State of the Science of Endocrine Disrupting Chemicals - 2012. Available: <http://www.unep.org/chemicalsandwaste/UNEPsWork/EndocrineDisruptingChemicals/tabid/130226/Default.aspx> [accessed 2 March 2015].
- Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, et al. 2008. ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr Comput Aided Drug Des* 4:191–198.
- Vedani A, Smiesko M. 2009. *In silico* toxicology in drug discovery—concepts based on three-dimensional models. *Altern Lab Anim* 37:477–496.
- Wedebeye EB, Niemelä JR, Nikolov NG, Dybdahl M, eds. 2013. Use of QSAR to Identify Potential CMR Substances of Relevance under the REACH Regulation. Environmental Project No. 1503. Copenhagen, Denmark:Danish Ministry of the Environment, Environmental Protection Agency. Available: <http://www2.mst.dk/Udgiv/publications/2013/09/978-87-93026-48-3.pdf> [accessed 27 May 2016].
- Wetmore BA, Wambaugh JF, Ferguson SS, Sochaski MA, Rotroff DM, Freeman K, et al. 2012. Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. *Toxicol Sci* 125:157–174.
- Wold S, Sjöström M, Eriksson L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometr Intell Lab Syst* 58:109–130.
- Worth AP, Bassan A, Gallegos A, Notzeva TI, Patlewicz G, Pavan M, et al. 2005. The Characterisation of (Quantitative) Structure–Activity Relationships: Preliminary Guidance. Ispra, Italy:European Commission Joint Research Centre. EUR 21866 EN. Available: https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/doc/QSAR_characterisation_EUR_21866_EN.pdf [accessed 27 May 2016].
- Xie Q, Ratnasinghe LD, Hong H, Perkins R, Tang ZZ, Hu N, et al. 2005. Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer: a novel method. *BMC Bioinformatics* 6(suppl 2):S4, doi: 10.1186/1471-2105-6-S2-S4.
- Yap CW. 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 32:1466–1474.
- Zakharov AV, Peach ML, Sitzmann M, Nicklaus MC. 2014. A new approach to radial basis function approximation and its application to QSAR. *J Chem Inf Model* 54:713–719.
- Zang Q, Rotroff DM, Judson RS. 2013. Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure–activity relationship and machine learning methods. *J Chem Inf Model* 53:3244–3261.

3.4.4 My Further Remarks to CERAPP

The approach applied in CERAPP has its limitations both with regard to the biological endpoint and the methods for evaluating the individual models and constructing the consensus model. First, the U.S. EPA NCCT provided ToxCast training sets was derived from a network model that integrates results from 18 *in vitro* assays [18]. These 18 assays covers the steps of the classical ER signaling pathway starting from ligand binding to the ER ligand binding domain, dimerization, co-factor recruitment and DNA binding as well as protein production and ER-induced proliferation for the ER agonists [18]. EDCs can affect estrogen signaling through other estrogen signaling pathways and indirect mechanisms [19–22]. Therefore the negative predictions from CERAPP should not be used for acquitting chemicals as having estrogen modulating potential.

The evaluation method used in CERAPP does not constitute a proper external validation of the models (section 2.3.1) as the evaluation set contains both U.S. EPA NCCT ToxCast training set structures and structures applied in other training sets. Thus, depending on the degree to which the evaluation set structures were also included in the training set of the models, the performance results are likely to be affected. The models with a high overlap of training and evaluation set structures have most likely also performed better in the evaluation. As described in the paper, the results from the evaluations were included in the assignment of the two model scores. These scores were subsequently used when constructing the consensus model. The potential bias introduced to these scores evaluations could hereby have influenced the constructed consensus model and its predictions. However, the reason for making the consensus model was to overcome the limitations of the single models in terms of their coverage and applied algorithms, and this was not compromised by the evaluation procedure. Also, the main goal of CERAPP was to use the consensus model predictions for prioritizing chemicals for further testing in EDSP and not to develop a high performance consensus model [3]. Performing true robust external validations of the many models included in CERAPP would have been both impractical and very time-consuming.

3.4.5 Conclusions

To conclude, the approach and predictions from CERAPP serve as useful prioritization tools for further testing of e.g. the EDSP universe, but the negative predictions cannot be used for classifying chemicals as non-EDCs just as the model evaluation results should not be interpreted as external validations. To conclude on the additional work made, the ER agonist model developed for CERAPP showed high predictive performance in an in-house robust cross-validation with balanced accuracy of 87.5%. In the screening of the REACH-PRS set the model could predict 73.7% of the substances and of these 4,198 chemicals were predicted as potential ER agonists.

References

- [1] EDSP, Federal Register: Part II Environmental Protection Agency - Endocrine Disruptor Screening Program: Statement of Policy; Notice, Priority-Setting Workshop; Notice (1998). <https://www.epa.gov/sites/production/files/2015-08/documents/122898frnotice.pdf> (accessed March 16, 2017).
- [2] EDSP, Federal Register: Environmental Protection Agency - Endocrine Disruptor Screening Program Notice (1998). <https://www.epa.gov/sites/production/files/2015-08/documents/081198frnotice.pdf> (accessed March 16, 2017).
- [3] US-EPA NCCT, CERAPP -Collaborative Estrogen Receptor Activity Prediction Project, (2016). <https://www.epa.gov/chemical-research/cerapp-collaborative-estrogen-receptor-activity-prediction-project-0> (accessed March 16, 2017).
- [4] EDSP, Federal Register: Environmental Protection Agency - Endocrine Disruptor Screening Program (EDSP); Announcing the Availability of the Tier 1 Screening Battery and Related Test Guidelines; Notice (2009). <https://www.federalregister.gov/documents/2009/10/21/E9-25348/endocrine-disruptor-screening-program-edsp-announcing-the-availability-of-the-tier-1-screening> (accessed January 19, 2017).
- [5] C.E. Willett, P.L. Bishop, K.M. Sullivan, Application of an Integrated Testing Strategy to the U.S. EPA Endocrine Disruptor Screening Program, *Toxicol. Sci.* 123 (2011) 15–25. doi:10.1093/toxsci/kfr145.
- [6] EDSP21 Work Plan, The Incorporation of In Silico Models and In Vitro High Throughput Assays in the Endocrine Disruptor Screening Program (EDSP) for Prioritization and Screening, (2011). https://www.epa.gov/sites/production/files/2015-07/documents/edsp21_work_plan_summary_overview_final.pdf (accessed March 13, 2017).
- [7] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A.M. Richard, C.M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E.B. Wedebye, F. Grisoni, G.F. Mangiatordi, G.M. Incisivo, H. Hong, H.W. Ng, I. V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N.G. Nikolov, O. Nicolotti, P.L. Andersson, Q. Zang, R. Politi, R.D. Beger, R. Todeschini, R. Huang, S. Farag, S.A. Rosenberg, S. Slavov, X. Hu, R.S. Judson, CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ. Health Perspect.* 124 (2016) 1023–1033. doi:10.1289/ehp.1510267.
- [8] AOP-8, Upregulation of Thyroid Hormone Catabolism via Activation of Hepatic Nuclear Receptors, and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/8> (accessed March 13, 2017).
- [9] Y. Tsuchiya, M. Nakajima, T. Yokoi, Cytochrome P450-mediated metabolism of estrogens and its regulation in human, *Cancer Lett.* 227 (2005) 115–124. doi:10.1016/j.canlet.2004.10.007.
- [10] J.-M. Pascussi, S. Gerbal-Chaloin, L. Drocourt, E. Assénat, D. Larrey, L. Pichard-Garcia, M.-J. Vilarem, P. Maurel, Cross-talk between xenobiotic detoxication and other signalling pathways: clinical and toxicological consequences, *Xenobiotica*. 34 (2004) 633–664. doi:10.1080/00498250412331285454.
- [11] J.-M. Pascussi, S. Gerbal-Chaloin, C. Duret, M. Daujat-Chavanieu, M.-J. Vilarem, P. Maurel, The Tangle of Nuclear Receptors that Controls Xenobiotic Metabolism and Transport: Crosstalk and Consequences, *Annu. Rev. Pharmacol. Toxicol.* 48 (2008) 1–32. doi:10.1146/annurev.pharmtox.47.120505.105349.
- [12] A.P. Santin, T.W. Furlanetto, Role of Estrogen in Thyroid Function and Growth Regulation, *J. Thyroid Res.* 2011 (2011) 1–7. doi:10.4061/2011/875125.

-
- [13] J. Fishman, L. Hellman, B. Zumoff, T.F. Gallagher, Effect of Thyroid on Hydroxylation of Estrogen in Man, *J. Clin. Endocrinol. Metab.* 25 (1965) 365–368. doi:10.1210/jcem-25-3-365.
- [14] Y.S. Zhu, P.M. Yen, W.W. Chin, D.W. Pfaff, Estrogen and thyroid hormone interaction on regulation of gene expression., *Proc. Natl. Acad. Sci.* 93 (1996) 12587–12592. doi:10.1073/pnas.93.22.12587.
- [15] T.L. Dellovade, Y.S. Zhu, L. Krey, D.W. Pfaff, Thyroid hormone and estrogen interact to regulate behavior, *Proc. Natl. Acad. Sci.* 93 (1996) 12581–12586. doi:10.1073/pnas.93.22.12581.
- [16] QSARDB, Danish (Q)SAR Database, (2015). <http://qsar.food.dtu.dk/> (accessed March 14, 2017).
- [17] S.A. Rosenberg, M. Xia, R. Huang, N.G. Nikolov, E.B. Wedebye, M. Dybdahl, QSAR development and profiling of 72,524 REACH substances for PXR activation and CYP3A4 induction, *Comput. Toxicol.* 1 (2017) 39–48. doi:10.1016/j.comtox.2017.01.001.
- [18] R.S. Judson, F.M. Magpantay, V. Chickarmane, C. Haskell, N. Tania, J. Taylor, M. Xia, R. Huang, D.M. Rotroff, D.L. Filer, K.A. Houck, M.T. Martin, N. Sipes, A.M. Richard, K. Mansouri, R.W. Setzer, T.B. Knudsen, K.M. Crofton, R.S. Thomas, Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor, *Toxicol. Sci.* 148 (2015) 137–154. doi:10.1093/toxsci/kfv168.
- [19] N. Heldring, A. Pike, S. Andersson, J. Matthews, G. Cheng, J. Hartman, M. Tujague, A. Strom, E. Treuter, M. Warner, J.-Å. Gustafsson, Estrogen Receptors: How Do They Signal and What Are Their Targets, *Physiol. Rev.* 87 (2007) 905–931. doi:10.1152/physrev.00026.2006.
- [20] S. Nilsson, S. Mäkelä, E. Treuter, M. Tujague, J. Thomsen, G. Andersson, E. Enmark, K. Pettersson, M. Warner, J.A. Gustafsson, Mechanisms of estrogen action., *Physiol. Rev.* 81 (2001) 1535–1565.
- [21] E.R. Prossnitz, M. Barton, The G protein-coupled estrogen receptor GPER in health and disease, *Nat. Rev. Endocrinol.* 7 (2011) 715–726. doi:10.1038/nrendo.2011.122.
- [22] E.K. Shanle, W. Xu, Endocrine Disrupting Chemicals Targeting Estrogen Receptor Signaling: Identification and Mechanisms of Action, *Chem. Res. Toxicol.* 24 (2011) 6–19. doi:10.1021/tx100231n.

Part IV - In Closing

4.1 Overview

To recapitulate on the four projects in this thesis, a brief summary of each project and its main results is given below. The predictive performances of the QSAR models from each project as well as their coverages of the REACH-PRS set of 72,524 structure entries are summarized in Table 1.

Table 1. Overview of the predictive performances and coverage of the REACH-PRS set for the QSAR models developed in this thesis.

Overview		Cross-validation			External validation			REACH-PRS screening		
Project	QSAR models	Sens	Spec	BA	Sens	Spec	BA	Coverage (%)	POS_IN	NEG_IN
TPO	QSAR1	72.1	89.0	80.9	79.7	90.8	85.3	38,661 (53.3)	7,128	31,533
	QSAR2	75.6	89.8	82.7	-	-	-	45,540 (62.8)	8,790	36,750
PXR and CYP3A4	hPXR-LBD	68.7	84.5	76.6	85.0	87.8	86.4	43,551 (60.1)	11,490	32,061
	hPXR	72.5	80.4	76.4	80.0	85.2	82.6	38,114 (52.5)	6,167	31,947
	rPXR	58.9	92.0	75.4	91.3	94.1	92.7	52,144 (71.9)	3,141	49,003
	CYP3A4	71.6	80.7	76.1	76.9	85.5	81.2	42,861 (59.1)	5,874	36,987
AhR	QSAR4:1	-	-	-	85.1	97.2	91.2	46,261 (63.8)	1,269	44,992
	QSAR4:1-R	-	-	-	89.8	91.6	90.7	39,698 (54.7)	2,148	37,550
CERAPP	ER agonism	80.6	94.4	87.5	-	-	-	53,433 (73.7)	4,198	49,235

Sens = sensitivity, Spec = specificity, BA = balanced accuracy, AD = applicability domain, POS_IN = positive prediction in the defined AD, NEG_IN = negative predictions in the defined AD

Chapter 3.1: QSAR Models for TPO Inhibition *In Vitro*

The main aim of this project was to develop and apply global binary QSAR models for TPO inhibition, an important mechanism for thyroid disruption and an MIE in a thyroid-related AOP for DNT.

Main methods and results: Two QSAR models were built and validated:

- QSAR1: the training set consisted of 877 ToxCast phase I and II chemicals. The QSAR model underwent robust cross-validation as well as external validation with a large test set of 646 E1K ToxCast chemicals.
- QSAR2: the test set and training set for QSAR1 were merged to constitute a training set of 1,519 ToxCast chemicals, and a new larger QSAR model was built and cross-validated.

The cross-validation procedure was conservative compared to the external validation of QSAR1 (Table 1). Overall, both QSAR1 and QSAR2 showed high predictive performances according to their respective validations, i.e. balanced accuracies from 80.6% to 85.3% (Table 1). The top ten structural features in QSAR2 associated with TPO inhibition and non-inhibition, respectively, were identified,

and among structural features associated with TPO inhibition were versions of phenols, aniline and anisole. The EU REACH-PRS inventory and a US-EPA inventory of 32,197 unique structures were screened through QSAR1 and QSAR2. QSAR2 had approximately 10% larger coverages of REACH-PRS and US-EPA, which was an expected effect of expanding the training set (Table 1). The two isomers of BHA, both included in the inventories and used as e.g. food antioxidants, were used in a case study to exemplify one use of QSAR predictions, i.e. how QSAR predictions can aid in elucidating a chemical's mode-of-action(s) in AOs and support results from *in vivo* studies. The project has been described in a manuscript ready for submission.

Chapter 3.2: QSAR Models for PXR Interaction and CYP3A4 Induction *In Vitro*

The main aim of this project was to develop global binary QSAR models for PXR binding and activation as well as CYP3A4 induction. PXR regulates the expression of metabolizing enzymes, including CYP3A4, and some of these enzymes are involved in thyroid and estrogen hormone catabolism. PXR also regulates expression of proteins important for thyroid hormone membrane transport. Activation of PXR by xenobiotics can therefore induce thyroid disruption and is included as an MIE in an AOP for thyroid-related DNT.

Main methods and results: Four global binary QSAR models for hPXR-LBD binding, hPXR activation, rPXR activation and CYP3A4 induction, respectively, were built and underwent robust cross- and external validations. They were all robust and predictive with balanced accuracies of 75.4% to 76.6% in cross-validations and 82.6% to 92.7% in external validations (Table 1). The models were subsequently used for screening the REACH-PRS inventory, and could produce reliable predictions for 52.5% (hPXR) to 71.9% (rPXR) of the structures (Table 1). Concordance rates between relevant model endpoints were calculated on both the REACH-PRS predictions and the experimental data. From this, we saw a high overlap of 81% between predicted hPXR activators that were also predicted hPXR-LBD binders as well as between predicted hPXR activators being CYP3A4 inducers (88.4%) and vice versa (97.5%). We did not see any positive correlations between hPXR and rPXR activators, and these results emphasize the need to be careful when extrapolating rat toxicity data to humans. The project results have been published in [1] as an open access paper.

Chapter 3.3: QSAR Models for AhR Activation *In Vitro*

The main aim of this project was to use a large and highly imbalanced PubChem dataset for AhR activation to explore how a rational two-step selection of inactives for training set expansion would affect QSAR coverage and predictive performance. AhR, like PXR, regulates the expression of enzymes involved in estrogen and thyroid hormone catabolism, and AhR interaction is an MIE in a thyroid-related AOP for DNT.

Main methods and results: The large and imbalanced curated dataset was randomly split into a test set (93 actives and 154,513 inactives) and a dataset (832 actives and 50,000 inactives) for training set construction. The 832 training set actives were used in all training sets and different proportions of inactives were selected from the 50K set of inactives using two different approaches: random vs two-step rational selection. Two final QSAR models with an inactive to active ratio of 4:1 were made:

- QSAR4:1-R: consisted of the 832 actives and 3,328 inactives selected randomly from the 50K inactives.
- QSAR4:1: consisted of the 832 actives and 3,328 inactives selected in one random and two rational selection steps using predictions of the remaining 50K set structures in two intermediate models. This rational selection aimed at identifying and adding structures that could help expand the chemical space covered by the training set and improve the model's ability to correctly discriminate between actives and inactives.

The models were externally validated with the test set, and QSAR4:1 produced a higher number of true negative predictions and a smaller number of both false and true positive predictions compared to QSAR4:1-R. Thus, QSAR4:1 had a higher specificity (97.2% versus 91.6%) than QSAR4:1-R but a lower sensitivity (85.1% versus 89.8%) (Table 1). These results indicate that the two-step rational selection of inactives for QSAR4:1 has resulted in a model with an optimized ability to produce more reliable predictions of inactives at the expense of both correct and wrong active predictions. Then the models were used for screening of the REACH-PRS inventory. QSAR4:1 had around 9% larger coverage of the REACH-PRS set than QSAR4:1-R, i.e. 63.8% versus 54.7% (Table 1). For unknown reasons the same effect in coverages of the test set was not observed.

The projects in chapter 3.1, 3.2 and 3.3 cover relevant thyroid-related mechanisms and were all part of a project partly supported by a grant from the Danish 3R Center²³.

Chapter 3.4: The Collaborative Estrogen Receptor Activity Prediction Project

This project was part of the large international collaboration, CERAPP, organized by the U.S. EPA NCCT on building QSARs for the classical ER signaling pathway and using them to make consensus predictions for a CERAPP prediction set of around 32,500 U.S. EPA curated environmental chemicals. The output from CERAPP has been published in [2]. Activation of ER is an important mechanism in the endocrine system and is one of the best-studied effects of ECDs. It is indirectly related to thyroid hormone disruption due to e.g. ER cross-talk with thyroid-related mechanisms such as the AhR.

²³ <http://en.3rcenter.dk/research/projects/projects-2016/development-of-mechanism-based-computer-models-for-hazard-assessment-of-thyroid-hormone-disruption/>

Main methods and results: My contributions to CERAPP consisted of the development of a binary global QSAR model for ER agonism using a U.S. EPA provided training set. The model was rigorously cross-validated and showed high predictive performance in the cross-validation with a balanced accuracy of 87.5% (Table 1). The model and cross-validation were described in the QMRF format (Appendix), which was sent to U.S. EPA together with predictions of the CERAPP prediction set generated by the DTU Food QSAR team. U.S. EPA NCCT scientists performed evaluations of the individual models using an evaluation set included in the prediction set and used these results when they combined the corresponding model predictions provided by all the collaborators to reach consensus predictions on the CERAPP prediction set. The U.S. EPA evaluation set was not screened for training set overlap and could therefore not be used for external validation but only to weigh the single model predictions in the CERAPP consensus prediction. Besides the work made for CERAPP, I also applied the ER agonism QSAR model to screen the REACH-PRS set, and the model could make reliable predictions for 53,433 (73.7%) of the structures, and of these 4,198 were predicted ER agonists (Table 1).

Each project has been discussed in the respective project chapters. The next chapter contains a more general discussion of all four projects in relation to the background chapters followed by some concluding remarks and a short reflection on future research perspectives.

4.2 Discussion

The thyroid-relevant mechanisms covered in the projects of the PhD thesis include inhibition of TPO and interaction with the two NRs, PXR and AhR. The selection of these mechanisms for global binary QSAR development was primarily based on the availability of large and structurally diverse datasets with high quality experimental results as well as their relevance in established thyroid-related AOPs for DNT. Also, the selected datasets had to be useful for QSAR modeling, i.e. they should have contained sufficient data for both activity classes. The inclusion of the CERAPP project (3.4) on ER agonism in the PhD project was mainly due to the invitation from the U.S. EPA NCCT to participate. Such participation was a great opportunity to strengthen the collaboration with the U.S. EPA NCCT for future QSAR development projects.

4.2.1 Collection, Curation and Preparation of the Applied Datasets

The training and validation sets in each project were collected from the same sources, respectively, and the experimental data had undergone the same testing protocol(s) and data analysis. Furthermore, in project 3.1 and 3.2 the models were developed in close collaboration with the data providers. In all the projects, the chemical structures underwent a structure curation procedure to remove structures unacceptable for QSAR processing. Most assays are associated with artefacts

related to the applied technology, e.g. luciferase or fluorescence interference, or protocol, e.g. cytotoxicity in cell cultures. Such artefacts can result in false positive or negative experimental results [3]. In the curation procedure of the datasets for 3.1, 3.3 and 3.4, different steps were taken to identify such potentially false experimental results. In 3.1 and 3.4, the U.S. EPA NCCT provided data had previously undergone different curation procedures using information from related assays to flag potentially false experimental results. For the AhR project, available PubChem data for luciferase interference were used as a counterscreen to flag potential false active results. Based on the flags for potential assay interference, we classified portions of the data entries as inconclusive for the given endpoint and excluded them from the subsequent model development. The structure curation and exclusion of inconclusive and potentially false experimental results have contributed to reducing the noise in the datasets.

4.2.2 QSAR Development

All the training and test sets were large and diverse enough to build global QSAR models and perform large external validations, respectively. Only QSAR models with binary, i.e. active versus inactive, response variables were made in this PhD project. This was done mainly due to the nature of the provided data. None of the models have had any outliers removed, and thus all available information to the extent possible was used in the model development. Wherever possible, the built QSAR models underwent both large external validation and rigorous five times two-fold cross-validation to assess their predictive performances in the defined AD (Table 1). The experience from project 3.1 and 3.2 was that the applied cross-validation procedure underestimates the predictive performance compared to applied large external validations. Goodness-of fit tests have not been performed in the projects but have been made subsequently, and the results are available in Table 2.

Table 2. Goodness-of-fit results of the QSAR models developed in this thesis.

Goodness-of-fit		Predictions in AD				Statistical parameters		
Project	Models	TP	FP	TN	FN	Sensitivity	Specificity	Balanced accuracy
TPO	QSAR1	84	37	491	2	97.7	93.0	95.4
	QSAR2	147	53	846	13	91.3	94.1	92.7
PXR and CYP3A4	hPXR-LBD	111	117	892	6	94.9	88.4	91.7
	hPXR	133	120	757	11	92.4	86.3	89.4
	rPXR	81	65	1214	3	96.4	94.9	95.7
	CYP3A4	127	173	865	11	92.0	83.3	87.7
AhR	QSAR4:1	466	140	1965	37	92.6	93.3	93.0
	QSAR4:1-R	591	157	2475	42	93.4	94.0	93.7
CERAPP	ER agonism	64	52	1090	5	92.8	95.4	94.1

TP = true positive, FP = false positive, TN = true negative, FN = false negative, AD = applicability domain

As expected, when comparing the balanced accuracies from the external and/or cross-validations (Table 1) with the corresponding goodness-of-fit balanced accuracies (Table 2), the goodness-of-fit results were better in all cases. Since all models showed good predictive performances with balanced accuracies over 75% in the cross-validations and 82% in the external validations (Table 1) this indicates that the models are able to generalize and have not been overfitted to their training sets.

The good predictive performances of the models are likely a result of a combination of the following:

- An overall high quality of the experimental datasets including the fact that all data in the respective datasets originated from the same source with experimental results from the same test protocol(s)
- The structure and data curation steps to reduce noise in the datasets
- The use of the composite model function in LPDM to increase performance of the smaller class in the imbalances training sets, i.e. sensitivity in these cases
- The chemical descriptors and modeling method were adequate for the modeled endpoints
- The application of a 'strict' AD to exclude the likely less reliable predictions from the statistical analyses

4.2.3 Limitations of the Developed QSAR Models

QSARs are, like other *in silico*, *in vitro* or *in vivo* studies, models that serve to estimate the true values, and false predictions are in general an unavoidable attribute of any (QSAR) model [4]. Validation of a model can provide measures of how good the model is at making correct estimates and information about the uncertainty in these estimates. As QSAR models are trained on experimental data from *in vitro* or *in vivo* models their predictive performance depend on the performance of the underlying experimental data. In theory a model can be more precise than the experimental results, but this is rare and difficult to prove. False predictions produced from the QSAR models can be a result of wrong information included in the model, e.g. due to unforeseen artefacts in the experimental data model or unknown chemical impurities causing the activity. They may also be due to the more rare cases where the QSAR, with help from its knowledge from training set structural analogs, have identified a wrong experimental result. Furthermore, a false QSAR prediction may reflect that the underlying similarity hypothesis is not bullet-proof, for example due to 'activity cliffs', i.e. areas in the chemical space where a small change in the chemical structure can have a dramatic effect on its activity [5–7]. If such information have not been included in the training of the model, then the model is unlikely to be able to identify such 'activity cliffs' when applied on

new structures. Finally, wrong predictions may be due to inappropriateness of the used modeling method or descriptors, as well as other reasons.

The results from the robust cross- and external validation studies of the QSAR models described in this thesis gives useful information to the model user. The sensitivity and specificity measures quantify how good a model is at avoiding false negative and false positive predictions, respectively. For any test there is usually a trade-off between these two measures and whether a high specificity or a high sensitivity is preferred depends on the purpose of the model. If the purpose is to identify as many positives as possible and avoid false negative predictions then a model with a high sensitivity is preferable, however at the expense of risking a high rate of false positives. If the purpose is to be quite certain that a positive prediction is correct then a model with high specificity would be preferred. All models in this thesis had higher specificity than sensitivity in their validation(s) (Table 1). This was mainly an effect of the higher ratio of inactives in the training sets but also partly driven by a deliberate choice in the modeling procedures

4.2.4 Using the Developed QSAR Models

The QSAR models developed in the PhD project can serve multiple uses and some have already been mentioned in the project chapters. Here a few examples are given and discussed in terms of their use limitations.

For Screening and Prioritization

Global QSAR models are useful tools for virtual screening of large chemical libraries. In the present PhD project, the developed global QSAR models were among other things applied to screen the large chemical inventory of 72,524 REACH-PRS substances. The models could predict between 38,114 (52.5%) to 53,433 (73.7%) of the REACH-PRS structures in their respective ADs (Table 1). In this way the developed global QSAR models succeeded to substantially expand the experimental knowledge from the 1,000s of chemical structures they were trained on, and the QSAR-derived information on 10,000s of chemicals can contribute to the identification and prioritization of potential EDCs, mainly TDCs, for further evaluations. As the models have high specificities we expect a fairly high rate of true positives among the positive predictions from the screenings but also a relatively high risk of not catching some positives due to many false negative predictions. Corresponding predictions from the developed models, as well as previously built QSARs, can also be used in combination to identify chemicals that are both inhibiting TH synthesis, i.e. are TPO inhibitors, and increasing TH catabolism, e.g. through PXR and/or AhR activation. Chemicals that affect both TH synthesis and catabolism are likely to have a more pronounced effect on TH levels and could be ranked as the highest priority chemicals. As all of the models have been trained to predict binary endpoints they cannot output

information of the chemicals potencies for the given mechanism. Such information could also have been useful in a ranking.

In Research

The QSAR models may aid in the development, optimization or repurposing of chemicals and drugs, for example drugs for treatment of thyroid-related diseases. They may also be used for generating new hypotheses on molecular mechanisms in AOs by searching for statistical correlations between chemicals predicted active for e.g. TPO inhibition and having data for an AO. Such data-driven associations will have to be investigated further in animal models to be confirmed or rejected. Finally, predictions from the present models can aid in the design of *in vivo* toxicity studies of chemicals by providing information on the chemical's possible mode-of actions and potential AOs that could be investigated.

In Regulatory Contexts

Whether the developed models are applicable for regulatory use does not only depend on their ability to provide reliable predictions, but also of their regulatory relevance [8]. The developed models from the present project are of regulatory relevance and may serve multiple applications in regulatory contexts. They can for example provide information to fill datagaps or aid in groupings and read-across cases (see e.g. [9]). While predictions from the developed QSARs can be used to raise suspicion that a chemical may cause an AO, they are not on their own sufficient to definitively assess this. For this purpose, they should be used e.g. in combination with relevant AOPs, and together this information can feed into an IATA on chemical assessment. The QSAR models are all based on data from *in vitro* studies and it is therefore important to also include information of a chemical's toxicokinetics in the assessment [10]. The guidance document for triggers of the EOGRTS DNT cohort inclusion under REACH is still under development [11], and, depending on its outcome, it is likely that the QSAR models in combination with relevant DNT AOP(s) can be included in future triggers for DNT testing in EOGRTS.

4.3 Concluding Remarks

The validation studies show that the developed global QSAR models for the selected MIEs of thyroid-related AOPs and the ER agonism model are robust and highly predictive. The application of the models to predict large inventories containing 10,000s of man-made chemicals showed that these global models are able to generate reliable predictions for more than half of the chemicals in the inventories. In this way, the models were able to greatly expand the knowledge derived from experimental data on thousands of chemicals to provide prediction information on tens of

thousands of untested chemical structures for their potential interaction with MIEs in relevant AOPs. The QSAR models of this thesis can in this way aid in the human safety evaluation of chemicals.

4.4 Perspectives

All the models developed in this PhD projects will be used for screening a structure set of more than 640,000 structures, and the predictions will be made freely available in the online Danish (Q)SAR Database [12]. Furthermore, the models will also been made available in a free, online QSAR model website (under construction), where they can be applied to predict the activity of user-submitted structures. If additional and adequate experimental data for the modeled MIEs become available, this can possibly in the future be used for further validation studies of the models and/or merged with the existing training sets to build larger QSARs with enhanced ADs that possibly can predict larger portions of the chemical universe.

The QSAR models in this PhD project only cover a few of the mechanisms in the thyroid system and other mechanisms not covered in the present PhD project include inhibition of NIS or deiodinases, interaction with TTR, TBG, TRs or TSH receptor as well as interaction with membrane transport proteins [10]. For most of these mechanisms there were either not (enough) experimental data available during the course of the PhD, e.g. NIS inhibition, or the available datasets were assessed sub-optimal for global QSAR development, for example due to too few known actives, e.g. for TR binding [10]. Time was of course also a limiting factor for not including more mechanisms in the project. Efforts to develop and apply HTS assay for other relevant mechanisms in thyroid/endocrine disruption is ongoing [10,13,14]. Examples on thyroid-relevant HTS data underway include data for NIS [15] and deiodinase inhibition [16], and the data could be used for future QSAR modeling studies. A battery of global QSAR models for a range of relevant thyroid/endocrine mechanisms including those developed in this PhD and new QSARs will be of high value. In the (far) future such a battery of QSARs for MIEs and KEs together with relevant AOPs might replace traditional animal studies in regulatory toxicology.

References

- [1] S.A. Rosenberg, M. Xia, R. Huang, N.G. Nikolov, E.B. Wedebye, M. Dybdahl, QSAR development and profiling of 72,524 REACH substances for PXR activation and CYP3A4 induction, *Comput. Toxicol.* 1 (2017) 39–48. doi:10.1016/j.comtox.2017.01.001.
- [2] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A.M. Richard, C.M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E.B. Wedebye, F. Grisoni, G.F. Mangiatordi, G.M. Incisivo, H. Hong, H.W. Ng, I. V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N.G. Nikolov, O. Nicolotti, P.L. Andersson, Q. Zang, R. Politi, R.D. Beger, R. Todeschini, R. Huang, S. Farag, S.A. Rosenberg, S. Slavov, X. Hu, R.S. Judson, CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ. Health Perspect.* 124 (2016) 1023–1033. doi:10.1289/ehp.1510267.
- [3] R. Judson, R. Kavlock, M. Martin, D. Reif, K. Houck, T. Knudsen, A. Richard, R.R. Tice, M. Whelan, M. Xia, R. Huang, C. Austin, G. Daston, T. Hartung, J.R. Fowle III, W. Wooge, W. Tong, D. Dix, Perspectives on validation of high-throughput assays supporting 21st century toxicity testing, *ALTEX.* 30 (2013) 51–56. doi:10.14573/altex.2013.1.051.
- [4] G.E.P. Box, Science and Statistics, *J. Am. Stat. Assoc.* 71 (1976) 791–799. doi:10.2307/2286841.
- [5] K.P. Cross, R.D. Benz, L. Stavitskaya, N.L. Kruhlak, Identifying Structure-Activity Cliffs in a Salmonella QSAR Model for Predicting the Potential Mutagenicity of Genotoxic Drug Impurities and Other Organic Molecules, (2012). http://www.leadscope.com/media/EMS_2012-IdentifyingStructureActivityCliffs.pdf (accessed March 27, 2017).
- [6] G.M. Maggiora, On Outliers and Activity Cliffs - Why QSAR Often Disappoints, *J. Chem. Inf. Model.* 46 (2006) 1535–1535. doi:10.1021/ci060117s.
- [7] A. Tropsha, Best Practices for QSAR Model Development, Validation, and Exploitation, *Mol. Inform.* 29 (2010) 476–488. doi:10.1002/minf.201000061.
- [8] ECHA, Guidance on information requirements and chemical safety assessment - Chapter R.6: QSARs and grouping of chemicals, (2008). https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf (accessed March 16, 2017).
- [9] Danish EPA, Category approach for selected brominated flame retardants, (2016). <http://www2.mst.dk/Udgiv/publications/2016/07/978-87-93435-90-2.pdf> (accessed February 17, 2017).
- [10] A.J. Murk, E. Rijntjes, B.J. Blaauboer, R. Clewell, K.M. Crofton, M.M.L. Dingemans, J. David Furlow, R. Kavlock, J. Köhrle, R. Opitz, T. Traas, T.J. Visser, M. Xia, A.C. Gutleb, Mechanism-based testing strategy using in vitro approaches for identification of thyroid hormone disrupting chemicals, *Toxicol. Vitr.* 27 (2013) 1320–1346. doi:10.1016/j.tiv.2013.02.012.
- [11] EC, Commission Regulation (EU) 2015/282 of 20 February 2015 amending Annexes VIII, IX and X to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as regards the Extended One-Generation Reproductive Toxicity Study, (2015). <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R0282&rid=1>.
- [12] QSARDB, Danish (Q)SAR Database, (2015). <http://qsar.food.dtu.dk/> (accessed March 14, 2017).
- [13] OECD, New scoping document on in vitro and ex vivo assays for the identification of

- modulators of thyroid hormone signalling, *OECD Environ. Heal. Saf. Publ.* . (2014).
[http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2014\)23&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2014)23&doclanguage=en) (accessed March 13, 2017).
- [14] K. Paul Friedman, S. Papineni, M.S. Marty, K.D. Yi, A.K. Goetz, R.J. Rasoulpour, P. Kwiatkowski, D.C. Wolf, A.M. Blacker, R.C. Pepper, A predictive data-driven framework for endocrine prioritization: a triazole fungicide case study, *Crit. Rev. Toxicol.* 46 (2016) 785–833. doi:10.1080/10408444.2016.1193722.
- [15] D.R. Hallinger, A.S. Murr, A.R. Buckalew, S.O. Simmons, T.E. Stoker, S.C. Laws, Development of a screening approach to detect thyroid disrupting chemicals that inhibit the human sodium iodide symporter (NIS), *Toxicol. Vitr.* 40 (2017) 66–78. doi:10.1016/j.tiv.2016.12.006.
- [16] U.S. EPA, Screening the ToxCast Phase I Chemical Library for inhibition of Deiodinase Type I enzyme activity, (2017).
https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=335810 (accessed March 25, 2017).

Appendix

QMRF: Model for mammalian Estrogen Receptor agonism *in vitro* (CERAPP)

1. QSAR identifier

1.1 QSAR identifier (title)

Leadscope Enterprise model for the U.S. EPA overall conclusion regarding mammalian Estrogen Receptor agonism *in vitro* (CERAPP), model made by the Danish QSAR Group at DTU Food.

1.2 Other related models

No

2. General information

2.1 Date of QMRF

June 2014.

2.2 QMRF author(s) and contact details

QSAR Group at DTU Food;

Danish National Food Institute at the Technical University of Denmark;

<http://qsar.food.dtu.dk/>;

qsar@food.dtu.dk

Sine Abildgaard Rosenberg;

National Food Institute at the Technical University of Denmark;

siro@food.dtu.dk

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

ebawe@food.dtu.dk

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

nign@food.dtu.dk

Marianne Dybdahl;

National Food Institute at the Technical University of Denmark;

mdyb@food.dtu.dk

2.3 Date of QMRF update(s)

April 2017.

2.4 QMRF update(s)

1

2.5 Model developer(s) and contact details

Sine Abildgaard Rosenberg;

National Food Institute at the Technical University of Denmark;

siro@food.dtu.dk

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

ebawe@food.dtu.dk

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

nign@food.dtu.dk

Marianne Dybdahl;

National Food Institute at the Technical University of Denmark;

mdyb@food.dtu.dk

Danish QSAR Group at DTU Food;

National Food Institute at the Technical University of Denmark;

<http://qsar.food.dtu.dk/>;

qsar@food.dtu.dk

2.6 Date of model development and/or publication

June 2014.

2.7 Reference(s) to main scientific papers and/or software package

Roberts, G., Myatt, G. J., Johnson, W. P., Cross, K. P., and Blower, P. E. J. (2000) LeadScope: Software for Exploring Large Sets of Screening Data. *Chem. Inf. Comput. Sci.*, 40, 1302-1314. doi: 10.1021/ci0000631

Cross, K.P., Myatt, G., Yang, C., Fligner, M.A., Verducci, J.S., and Blower, P.E. Jr. (2003) Finding Discriminating Structural Features by Reassembling Common Building Blocks. *J. Med. Chem.*, 46, 4770-4775. doi:10.1021/jm0302703

Valerio, L. G., Yang, C., Arvidson, K. B., and Kruhlak, N. L. (2010) A structural feature-based computational approach for toxicology predictions. *Expert Opin. Drug Metab. Toxicol.*, 6:4, 505-518. doi: 10.1517/17425250903499286

2.8 Availability of information about the model

The training set was kindly provided by the U.S. Environmental Protection Agency (EPA) and is non-proprietary. The model algorithm is proprietary from commercial software. This model was made for the U.S. EPA CERAPP project.

3. Defining the endpoint

3.1 Species

Bovine, mouse and human cell lines (18 biochemical and cell-based *in vitro* assays).

3.2 Endpoint

QMRF 4. Human Health Effects

QMRF 4.18.b. Receptor binding and gene expression (Estrogen Receptor)

3.3 Comment on endpoint

There is increasing evidence that a variety of environmental chemicals have the potential to disrupt the endocrine system by mimicking or inhibiting endogenous hormones such as estrogens and androgens. These endocrine disrupting chemicals (EDCs) may adversely affect development and/or reproductive function.

Natural estrogens are involved in the development and adult function of organs of the female genital tract, neuroendocrine tissues and the mammary glands; their role in reproduction spans from maintenance of the menstrual cycle to pregnancy and lactation. These effects are primarily mediated through the estrogen receptors (ERs), members of the nuclear receptor superfamily. When estrogen binds to the ER in the cytoplasm a receptor-hormone complex dimer is formed. This dimer translocates to the nucleus, where it recruits co-factors to form the active transcription factor (TF) complex. The active TF binds to the estrogen response element upstream to the target gene. This binding activates transcription of mRNA and subsequent translation to proteins that exert the hormone effects. Two isoforms of the ER exist in humans, alpha and beta, and both are widely expressed in different tissue types although there are some differences in their expression pattern. Exogenous compounds able to bind to and activate the ERs (i.e. ER agonists) have the ability to mimic natural estrogens and cause adverse effects to the reproductive system. Likewise, exogenous compounds that bind to the ERs without subsequent activation (i.e. ER antagonists) can potentially disturb the effect of the natural estrogens by blocking the receptors.

Results from 18 *in vitro* high-throughput screening assays that probe the ER signalling pathway in a mammalian system were integrated in a computational network model (Judson et al. 2014). The assays were a combination of biochemical and cell-based *in vitro* assays and probe perturbations of the ER pathway at multiple sites: receptor binding, receptor dimerization, DNA binding of the active transcription factor, gene transcription and changes in ER-induced cell growth kinetics. The network model uses activity patterns across the 18 *in vitro* assays to predict whether the chemical is an ER agonist, an ER antagonist, or instead is causing activity through narrow (technology-specific) or broad assay interference. For example, if a chemical is active in all of the assays in the ER agonism pathway of the network model a score for agonism is calculated as the AUC for the accumulated Hill model (based on the AC50 from the assays). If none or only parts of the assays in the ER agonist pathway are active, the chemical is a clear negative or is causing some form of assay interference (narrow or broad depending on which assays in the pathway that are active), respectively. These chemicals have an ER agonist score of 0 and are all assumed to be negative (Judson et al. 2014).

In order to make a classification model, compounds with an ER agonist score of 0 were defined as inactives and compounds with an AUC score of 0.1 or above were defined as an ER agonist.

3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

3.5 Dependent variable

Mammalian Estrogen Receptor agonist: positive or negative.

3.6 Experimental protocol

See S1, Appendix 1 in Judson *et al.* 2015.

3.7 Endpoint data quality and variability

The data is expected to be of high quality because of the integration of several assays to exclude false positives caused by narrow (technology-specific) or broad assay interference. Also, the variability in the data is expected to be low as for each assay all chemicals have been tested in the same laboratory and the process of assigning an ER agonist score using the network model (see 3.2) has been equal for all chemicals.

4. Defining the algorithm

4.1 Type of model

A categorical QSAR model based on structural features and numeric molecular descriptors.

4.2 Explicit algorithm

This is a categorical QSAR model made by use of partial logistic regression (PLR). Because of the imbalanced training set the “mother model” is a composite model consisting of ten submodels, using all the positives (80 chemicals) in each of these and different sub-sets of the negatives (see 4.5). The specific implementation is proprietary within the Leadscape software.

4.3 Descriptors in the model

structural features,

aLogP,

polar surface area,

number of hydrogen bond donors,

Lipinski score,

number of rotational bonds,

parent atom count,

parent molecular weight,

number of hydrogen bond acceptors

4.4 Descriptor selection

Leadscape Predictive Data Miner (LPDM) is a commercial software program for systematic sub-structural analysis of a compound using predefined structural features stored in a template library. The feature library contains approximately 27,000 structural features and the structural features chosen for the library are motivated by those typically found in small molecules: aromatics, heterocycles, spacer groups, simple substituents. Additionally, LPDM also calculates eight molecular descriptors for each structure: the octanol/water partition coefficient (alogP), hydrogen bond acceptors, hydrogen bond donors, Lipinski score, atom count, parent compound molecular weight, polar surface area and rotatable bonds. It is further possible to generate training set-dependent structural features (scaffold generation) and use these features in the model building process. Redundant features are removed and the remaining features are used in the model building. The default automatic feature selection process in LPDM selects the top 30% of the features according to X^2 -test for a binary variable, or the top and bottom 15% according to t -test for a continuous variable. LPDM treats numeric property data as ordinal categorical data. If the input data is continuous such as IC_{50} or cLogP data, the user can determine how values are assigned to categories: the number of categories and the cutoff values between categories. (Roberts et al. 2000).

4.5 Algorithm and descriptor generation

For descriptor generation see 4.4.

After selection of features the LPDM program performs partial least squares (PLS) regression for a continuous response variable, or partial logistic regression (PLR) for a binary response variable, to build a predictive model. By default LPDM performs leave-one-out or leave-groups-out (in the latter case, the user can specify any number of repetitions and percentage of structures left out) cross validation on the training set depending on the size of the training set.

In this model because of the categorical outcome in the response variable PLR was used to build the predictive model. Because of the unbalanced training set (i.e. 80 positives vs. 1342 negatives) ten submodels for smaller individual training sets consisting of the 80 positives and an equal number of negatives selected by random among the 1342 negatives were made. The descriptors for each of the ten submodels were automatically selected from the LPDM feature library based solely on the training set compounds used to build the individual submodel and was not affected by the training set chemicals in the composite "mother model". Therefore, a different number of descriptors (structural features and molecular descriptors) were selected and distributed on varying number of PLS factors for each submodel.

4.6 Software name and version for descriptor generation

Leadscape Predictive Data Miner, a component of Leadscape Enterprise version 3.1.1-10.

4.7 Descriptors/chemicals ratio

The model system uses molecular descriptors and structural features specific to a group of structurally related chemicals from the global training set. Therefore estimations of the number of used descriptors may be difficult. In general, we estimate that the models effectively use an order of magnitude less descriptors than numbers of chemicals in the training set when we set our domain definition where we weed out low probability active and inactive predictions (see 5.1).

5. Defining Applicability Domain

5.1 Description of the applicability domain of the model

For assessing if a test compound is within the applicability domain of a given model LPDM examines whether the test compound bears enough resemblance to the training set compounds used for building the model (i.e. structural domain analysis). This is done by calculating the distance between the test compound and all compounds in the training set (distance equals 1 - similarity). The similarity score is based on the Tanimoto method. The numbers of neighbors is defined as the numbers of compounds in the training set that have a distance ≤ 0.7 with respect to the test compound. The higher the number of neighbors the more reliable the prediction for the test compound. Statistics of the distances are also calculated. Effectively no predictions are made for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be generated.

In addition to the general LPDM structural domain definition the Danish QSAR group has applied a further requirement to the applicability domain of the model. Only predictions with probability (p) equal to or greater than 0.7 were accepted for actives. Predictions with p equal to or less than 0.3 were accepted for inactives. Predictions within the structural domain but with $p = [0.5;0.7[$ and $p =]0.3;0.5[$ were defined as positives out of applicability domain and negatives out of applicability domain, respectively. When these predictions were weeded out the performance increased at the expense of a reduced coverage.

5.2 Method used to assess the applicability domain

The system does not generate predictions for test compounds which are not in the structural domain or for which the molecular descriptors could not be generated.

Only predictions with probability equal to or greater than 0.7 were accepted for actives and predictions with probability equal to or less than 0.3 were accepted for inactives.

5.3 Software name and version for applicability domain assessment

Leadscope Predictive Data Miner (LPDM), a component of Leadscope Enterprise version 3.1.1-10.

5.4 Limits of applicability

The Danish QSAR group applies an overall definition of structures acceptable for QSAR processing which is applicable for all the in-house QSAR software, i.e. not only LPDM. According to this definition accepted structures are organic substances with an unambiguous structure, i.e. so-called discrete organics defined as: organic compounds with a defined two dimensional (2D) structure containing at least two carbon atoms, only certain atoms (H, Li, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and I), and not mixtures with two or more 'big components' when analyzed for ionic bonds (for a number of small known organic ions assumed not to affect toxicity the 'parent molecule' is accepted). Calculation 2D structures (SMILES and/or SDF) are generated by stripping off ions (of the accepted list given above). Thus, all the training set chemicals are used in their non-ionized form. See 5.1 for further applicability domain definition.

6. Internal validation

6.1 Availability of the training set

Yes

6.2 Available information for the training set

SMILES

6.3 Data for each descriptor variable for the training set

No

6.4 Data for the dependent variable for the training set

All

6.5 Other information about the training set

1422 compounds are in the training set: 80 positives and 1342 negatives.

6.6 Pre-processing of data before modeling

The results from the 18 ER in vitro assays were integrated using a network model and scores for ER agonism and ER antagonism were assigned to each chemical by US EPA (Judson et al. 2014). The ER agonist scores were categorized in order to make a categorical QSAR model. A cut off of 0.1 and above were set and chemicals in this category were defined as being ER agonists (80 chemicals). Chemicals with an ER agonist score of 0 were defined as not being ER agonists (1342 chemicals). The chemicals with an ER agonist score between 0 and 0.1 were excluded from the training set.

6.7 Statistics for goodness-of-fit

Not performed.

6.8 Robustness – Statistics obtained by leave-one-out cross-validation

Not performed. (It is not a preferred measurement for evaluating large models).

6.9 Robustness – Statistics obtained by leave-many-out cross-validation

A five times two-fold cross-validation was performed. This was done by randomly removing 50% of the full training set used to make the “mother model”, where the 50% contains the same ratio of positive and negatives as the full training set. A new model (validation submodel) was created on the remaining 50% using the same settings in LPDM but with no information from the “mother model” regarding descriptor selection etc. The validation submodel was applied to predict the removed 50% (within the defined applicability domain). Likewise, a validation submodel was made on the removed 50% of the training set and this model was used to predict the other 50% (within the defined applicability domain). This was repeated five times.

Predictions from the ten submodels were pooled and Coopers statistics for the composite “mother model” were calculated. This gave the following results for the 74,0% ($5263 \cdot 100\% / (5 \cdot 1422)$) of the predictions which were within the applicability domains of the respective sub-models:

- Sensitivity (true positives / (true positives + false negatives)): $270 / (270 + 65) = 80.60\%$
- Specificity (true negatives / (true negatives + false positives)): $4650 / (4650 + 278) = 94.36\%$
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): $4920 / 5263 = 93.48\%$
- Balanced accuracy ((Sensitivity + specificity)/2): $(80.6\% + 94.36\%) / 2 = 87.5\%$

6.10 Robustness - Statistics obtained by Y-scrambling

Not performed.

6.11 Robustness - Statistics obtained by bootstrap

Not performed.

6.12 Robustness - Statistics obtained by other methods

Not performed.

7. External validation

7.1 Availability of the external training set

7.2 Available information for the external training set

7.3 Data for each descriptor variable for the external training set

7.4 Data for the dependent variable for the external training set

7.5 Other information about the training set

7.6 Experimental design of test set

7.7 Predictivity – Statistics obtained by external validation

7.8 Predictivity – Assessment of the external validation set

7.9 Comments on the external validation of the model

External validation was not performed.

8. Mechanistic interpretation

8.1 Mechanistic basis of the model

The global model identifies structural features and molecular descriptors which in the model development was found to be statistically significant associated with effect. Many predictions may indicate modes of action that are obvious for persons with expert knowledge for the endpoint.

8.2 A priori or posteriori mechanistic interpretation

The identified structural features and molecular descriptors may provide basis for mechanistic interpretation.

8.3 Other information about the mechanistic interpretation

9. Miscellaneous information

9.1 Comments

The model can be used to predict if a chemical is an ER agonist (i.e. has an ER agonist score equal to or above 0.1) according to the network model based on the 18 ER pathway *in vitro* assays.

9.2 Bibliography

Judson, R.S., Magpantay, F.M., Chickarmane, V., Haskell, C., Tania, N., Taylor, J., Xia, M., Huang, R., Rotroff, D.M., Filer, D.L., Houck, K.A., Martin, M.T., Sipes, N., Richard, A.M., Mansouri, K., Setzer, R.W., Knudsen, T.B., Crofton, K.M., and Thomas, R.S. (2015) Integrated Model of Chemical

Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. *Toxicol.Sci.*, 148, 137-154. doi:10.1093/toxsci/kfv168

9.3 Supporting information